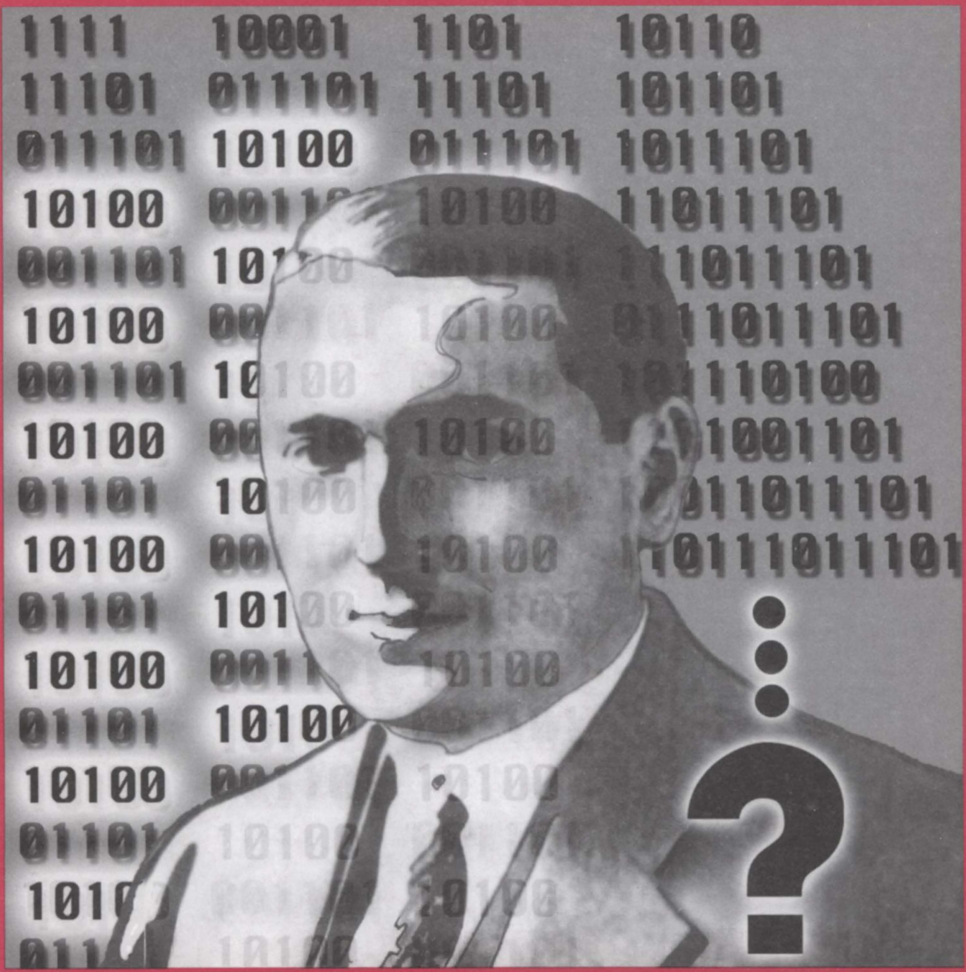
Vol. 77, No. 1, February 2004

MATHEMATICS MAGAZINE



Emil Post Ponders Tag Systems

- Emil Post and His Anticipation of Gödel and Turing
- Enter, Stage Center: The Early Drama of the Hyperbolic Functions
- Simpson Symmetrized and Surpassed

An Official Publication of The MATHEMATICAL ASSOCIATION OF AMERICA

EDITORIAL POLICY

Mathematics Magazine aims to provide lively and appealing mathematical exposition. The *Magazine* is not a research journal, so the terse style appropriate for such a journal (lemma-theorem-proof-corollary) is not appropriate for the *Magazine*. Articles should include examples, applications, historical background, and illustrations, where appropriate. They should be attractive and accessible to undergraduates and would, ideally, be helpful in supplementing undergraduate courses or in stimulating student investigations. Manuscripts on history are especially welcome, as are those showing relationships among various branches of mathematics and between mathematics and other disciplines.

A more detailed statement of author guidelines appears in this *Magazine*, Vol. 74, pp. 75–76, and is available from the Editor or at www.maa.org/pubs/mathmag.html. Manuscripts to be submitted should not be concurrently submitted to, accepted for publication by, or published by another journal or publisher.

Submit new manuscripts to Frank A. Farris, Editor, Mathematics Magazine, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0373. Manuscripts should be laser printed, with wide line spacing, and prepared in a style consistent with the format of *Mathematics Magazine*. Authors should mail three copies and keep one copy. In addition, authors should supply the full five-symbol 2000 Mathematics Subject Classification number, as described in *Mathematical Reviews*.

Cover image, *Emil Post Ponders Tag Systems*, by Brian Swimme. Number sequences stream down around Emil Post according to the tag system rules, described in Stillwell's article in this issue. The first three columns start with strings that lead to the same periodic pattern, but the future of the progression in the last column is unclear. Does it terminate, does it become periodic?

Cover image by Brian Swimme, a student at Santa Clara University, under the supervision of Jason Challas.

AUTHORS

John Stillwell is Professor of Mathematics at the University of San Francisco. He was born and raised in Melbourne, Australia, attending Melbourne High School and the University of Melbourne (M.Sc. 1965). He first came to America in 1965, and received Ph. D. in mathematics from MIT in 1970. At this time his main interests were in logic. From 1970 to 2001 he taught at Monash University in Melbourne, gradually changing fields to topology, geometry, algebra, and number theory. During this period he wrote a number of books, the best known of which is *Mathematics and Its History* (Springer-Verlag 1989, 2nd edition 2001). In 2002 he started a new career at the University of San Francisco.

Janet Heine Barnett holds a B.S. in mathematics and humanities from Colorado State University, and an M.A. and Ph.D. in set theory from the University of Colorado. She has taught at the University of Southern Colorado, where she has been a full professor since 1990. Her interests include mathematics history and its use in promoting mathematical understanding; her article in this MAGAZINE was inspired by questions from students who share these interests. Other projects include the mathematical history of Paris (jointly with her dance/travel partner/husband George Heine), and a study of proof and intuition in the development of mathematical concepts, both historical and in today's classroom.

Daniel J. Velleman received his B.A. from Dartmouth College in 1976 and his Ph.D. from the University of Wisconsin-Madison in 1980. He then taught at the University of Texas before joining the faculty of Amherst College in 1983. He is the author of *How to Prove It*, coauthor (with Stan Wagon and Joseph Konhauser) of *Which Way Did the Bicycle Go?*, and coauthor (with Alexander George) of *Philosophies of Mathematics*. His work on Simpson's rule was motivated by a nagging dissatisfaction, over a number of years of teaching Simpson's rule in calculus classes, with the alternation of coefficients in the Simpson's rule formula.

Vol. 77, No. 1, February 2004

MATHEMATICS MAGAZINE

MATHEMATICS MAGAZINE (ISSN 0025-570X) is published by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, D.C. 20036 and Montpelier, VT, bimonthly except July/August.
The annual subscription price for *MATHEMATICS MAGAZINE* to an individual member of the Association is $131. Student and unemployed members receive a 66% dues discount; emeritus members receive a 50% discount; and new members receive a 20% dues discount for the first two years of membership.)

Subscription correspondence and notice of change of address should be sent to the Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036. Microfilmed issues may be obtained from University Microfilms International, Serials Bid Coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

Advertising correspondence should be addressed to Dave Riska (*driska@maa.org*), Advertising Manager, the Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036.

Periodicals postage paid at Washington, D.C. and additional mailing offices.

Postmaster: Send address changes to Membership/ Subscriptions Department, Mathematical Association of America, 1529 Eighteenth Street, N.W., Washington, D.C. 20036-1385.

Printed in the United States of America

Emil Post and His Anticipation of Gödel and Turing

JOHN STILLWELL
University of San Francisco
San Francisco, CA 94117
stillwell@usfca.edu

Emil Post is known to specialists in mathematical logic for several ideas in logic and computability theory: the structure theory of recursively enumerable sets, degrees of unsolvability, and the Post "correspondence problem." However, he *should* be known to a much wider audience. In the 1920s he discovered the incompleteness and unsolvability theorems that later made Gödel and Turing famous. Post missed out on the credit because he failed to publish his results soon enough, or in enough detail. His achievements were known to most of his contemporaries in logic, but this was seldom acknowledged in print, and he now seems to be slipping into oblivion. Recent comprehensive publications, such as Gödel's collected works and the popular history of computation by Martin Davis [**3**] contain only a few words about Post, mostly in footnotes.

In this article I hope to redress the balance a little by telling Post's side of the story and presenting the gist of his ideas. This is not merely to give Post his due; it gives the opportunity to present Post's approach to Gödel's incompleteness theorem, which is not only more general than Gödel's but also simpler. As well as this, Post drew some nontechnical conclusions from the incompleteness theorem—about the interplay between symbolism, meaning, and understanding—that deserve wide circulation in mathematics classrooms.

Post's life and career

Post's life occupied roughly the first half of the 20th century. Here is a brief summary of the main events.

1897 February 11: born Augustów, Poland.

1904 May: emigrated to New York.

1917 B.S. from City College.

1920 Ph.D. from Columbia.

1921 Decidability and completeness of propositional logic in *Amer. J. Math.* Foresaw undecidability and incompleteness of general formal systems.

1936 Independent discovery of Turing machines in *J. Symb. Logic.*

1938 October 28: met with Gödel to outline his discoveries.

1941 Submitted his "Account of an Anticipation" to *Amer. J. Math.*

1944 Paper on recursively enumerable sets in *Bull. Amer. Math. Soc.*

1947 Proved unsolvability of word problem for semigroups in *J. Symb. Logic.*

1954 Died in New York.

I shall elaborate on his discoveries, particularly the unpublished ones, below. But first it is important to appreciate the personal background of his work. Post's life was in some ways a typical immigrant success story: His family brought him to New York as a child, he studied and worked hard and, with the help of a supportive wife and daughter, obtained a position at City College of New York and some renown in his field of research. However, life was tougher for Post than this brief outline would suggest.

When quite young he lost his left arm in an accident, and this ended his early dream of a career in astronomy. Around the age of 13, Post wrote to several observatories asking whether his disability would prevent his becoming an astronomer. Harvard College Observatory thought not, but the head of the U.S. Naval Observatory replied that it would, because "the use of both hands is necessary in all the work of this observatory." Post apparently took his cue from the latter, gave up on astronomy, and concentrated on mathematics instead.

He attended Townsend Harris High School and City College in New York, obtaining a B.S. in mathematics in 1917. As an undergraduate he did original work in analysis which was eventually published in 1930. It includes a result on the Laplace transform now known as the Post-Widder inversion formula. From 1917 to 1920, Post was a graduate student in mathematical logic at Columbia. Part of his thesis, in which he proves the completeness and consistency of the propositional calculus of Whitehead and Russell's *Principia Mathematica*, was published in the *American Journal of Mathematics* [**8**].

In 1920–1921 he held a post-doctoral fellowship at Princeton. During this time he tried to analyze the whole *Principia*, with a view to proving its completeness and consistency as he had done for propositional calculus. This was the most ambitious project possible, because the axioms of *Principia* were thought to imply all theorems of mathematics. Nevertheless, Post made some progress: He showed that all theorems of *Principia* (and probably of any conceivable symbolic logic) could be derived by simple systems of rules he called *normal systems*. At first this looked like a great step forward. But as he struggled to analyze even the simplest normal systems, Post realized that the situation was the opposite of what he had first thought: instead of simplifying *Principia*, he had merely distilled its complexity into a smaller system.

Sometime in 1921, as he later claimed, he caught a glimpse of the true situation:

- Normal systems can simulate any symbolic logic, indeed any mechanical system for deriving theorems.
- This means, however, that all such systems can be mechanically listed, and the diagonal argument then shows that the general problem of deciding whether a given theorem is produced by a given system is unsolvable.
- It follows, in turn, that no consistent mechanical system can produce all theorems.

I shall explain these discoveries of Post in more detail below. They include (in different form) the discoveries of Turing on the nature of computability and unsolvability, and Gödel's theorem on the incompleteness of formal systems for mathematics.

In 1921, Post suffered an attack of manic-depressive illness (as bipolar disorder was known at the time), and his work was disrupted at the height of his creative fever. The condition recurred quite frequently during his life, necessitating hospitalization and preventing Post from obtaining an academic job until 1935. To avert the manic episodes, Post would give himself two problems to work on, switching *off* the one that was going well when he found himself becoming too excited. This did not always work, however, and Post often received the electroshock treatment that was thought effective in those days. (His death from a heart attack at the early age of 57 occurred shortly after one such treatment.)

In 1935, Post gained a foothold in academia with a position at City College of New York. The teaching load was 16 hours per week, and all faculty shared a single large office, so Post did most of his research at home, where his daughter Phyllis was required not to disturb him and his wife Gertrude handled all day-to-day concerns. As Phyllis later wrote (quoted by Davis [**2**]):

> My father was a genius; my mother was a saint ... the buffer in daily life that permitted my father to devote his attention to mathematics (as well as to his varied interests in contemporary world affairs). Would he have accomplished so much without her? I, for one, don't think so.

By this time Post had seen two of his greatest ideas rediscovered by others. In 1931 Gödel published his incompleteness theorem, and in 1935 Church stated *Church's thesis*, which proposes a definition of computability and implies the existence of unsolvable problems. Church's definition of computability was not immediately convincing (at least not to Gödel), and some equivalent definitions were proposed soon after. The one that convinced Gödel was Turing's [**14**], now known as the *Turing machine*. Post's normal systems, another equivalent of the computability concept, were still unpublished. But this time Post had a little luck. Independently of Turing, and at the same time, he had reformulated his concept of computation—and had found a concept virtually identical with Turing's! It was published in a short paper [**9**] in the 1936 *Journal of Symbolic Logic*, with a note from Church affirming its independence from Turing's work.

This gave Post some recognition, but he was still in Turing's shadow. Turing had written a fuller paper, with clearer motivation and striking theorems on the existence of a universal machine and unsolvable problems. The world knew that Post had also found the definition of computation, but did not know that he had already seen the *consequences* of such a definition in 1921. In 1938, he met Gödel and tried to tell him his story. Perhaps the excitement was too much for Post, because he seems to have feared that he had not made a good impression. The next day, October 29, 1938, he sent Gödel a postcard that reads as follows:

> I am afraid that I took advantage of you on this, I hope but our first meeting. But for fifteen years I had carried around the thought of astounding the mathematical world with my unorthodox ideas, and meeting the man chiefly responsible for the vanishing of that dream rather carried me away.
>
> Since you seemed interested in my way of arriving at these new developments perhaps Church can show you a long letter I wrote to him about them. As for any claims I might make perhaps the best I can say is that I would have have *proved* Gödel's theorem in 1921—had I been Gödel.

After a couple more letters from Post, Gödel replied. He courteously assured Post that he had not regarded Post's claims as egotistical, and that he found Post's approach interesting, but he did not take the matter any further.

In 1941, Post made another attempt to tell his story, in a long and rambling paper "Absolutely unsolvable problems and relatively undecidable propositions—an account of an anticipation" submitted to the *American Journal of Mathematics*. The stream-of-consciousness style of parts of the paper and lack of formal detail made it unpublishable in such a journal, though Post received a sympathetic reply from the editor, Hermann Weyl. On March 2, 1942, Weyl wrote

> ... I have little doubt that twenty years ago your work, partly because of its revolutionary character, did not find its true recognition. However, we cannot turn the clock back ... and the American Journal is not the place for historical accounts ... (Personally, you may be comforted by the certainty that most of

the leading logicians, at least in this country, know in a general way of your anticipation.)

Despite these setbacks Post continued his research. In fact his most influential work was yet to come. In 1943, he was invited to address the American Mathematical Society, and his writeup of the talk [**11**] introduced his groundbreaking theory of *recursively enumerable sets*. Among other things, this paper sets out his approach to Gödel's theorem, which is perhaps ultimate in both simplicity and generality. This was followed in 1945 by a short paper [**12**], which introduces the "Post correspondence problem," an unsolvable problem with many applications in the theory of computation. The correspondence problem can be viewed as a problem about free semigroups, and in 1947, Post showed the unsolvability of an even more fundamental problem about semigroups—the *word problem* [**13**].

The unsolvability of this problem is the first link in a chain between logic and group theory and topology. The chain was completed by Novikov [**7**] in 1955, who proved the unsolvability of the word problem for groups, by Markov [**6**] in 1958, who deduced from it the unsolvability of the homeomorphism problem for compact manifolds, and by Higman [**5**] in 1961, who showed that "computability" in groups is equivalent to the classical concept of finite generation.

Thus Post should be celebrated, not only for his fundamental work in logic, but also for constructing a bridge between logic and classical mathematics. Few people today cross that bridge, but perhaps if Post's work were better known, more would be encouraged to make the journey.

Formal systems

In the late 19th century several new branches of mathematics emerged from problems in the foundations of algebra, geometry, and analysis. The rise of new algebraic systems, noneuclidean geometry, and with them the need for new foundations of analysis, created the demand for greater clarity in both the subject matter and methods of mathematics. This led to:

1. Symbolic logic, where all concepts of logic were expressed by symbols and deduction was reduced to the process of applying *rules of inference*.
2. Set theory, in which all mathematical concepts were defined in terms of *sets* and the relations of *membership* and *equality*.
3. Axiomatics, in which theorems in each branch of mathematics were deduced from appropriate *axioms*.

Around 1900, these branches merged in the concept of a *formal system*, a symbolic language capable of expressing all mathematical concepts, together with a set of propositions (axioms) from which theorems could be derived by specific rules of inference. The definitive formal system of the early 20th century was the *Principia Mathematica* of Whitehead and Russell [**15**].

The main aims of *Principia Mathematica* were *rigor* and *completeness*. The symbolic language, together with an explicit statement of all rules of inference, allows theorems to be derived only if they are logical consequences of the axioms. It is impossible for unconscious assumptions to sneak in by seeming "obvious." In fact, all deductions in the *Principia* system can in principle be carried out *without knowing the meaning of the symbols*, since the rules of inference are pure symbol manipulations. Such deductions can be carried out by a machine, although this was not the intention of *Principia*, since suitable machines did not exist when it was written. The intention was to ensure rigor by keeping out unconscious assumptions, and in these terms *Principia* was a complete success.

As for completeness, the three massive volumes of *Principia* were a "proof by intimidation" that all the mathematics then in existence was deducible from the *Principia* axioms, but no more than that. It was not actually known whether *Principia* was even *logically* complete, that is, capable of deriving all valid principles of logic. In 1930, Gödel proved its logical completeness, but soon after he proved its *mathematical* incompleteness. We are now getting ahead of our story, but the underlying reason for Gödel's incompleteness theorem can be stated here: the weakness of *Principia* (and all similar systems) is its very objectivity. Since *Principia* can be described with complete precision, *it is itself a mathematical object*, which can be reasoned about. A simple but ingenious argument then shows that *Principia* cannot prove all facts about itself, and hence it is mathematically incomplete.

Post's program

Post began his research in mathematical logic by proving the completeness and consistency of *propositional logic*. This logic has symbols for the words *or* and *not*—today the symbols $\vee$ and $\neg$ are commonly used—and variables $P, Q, R, \ldots$ for arbitrary propositions. For example, $P \vee Q$ denotes "P or Q", and $(\neg P) \vee Q$ denotes "(not P) or Q". The latter is commonly abbreviated $P \to Q$ because it is equivalent to "P implies Q".

Principia Mathematica gave certain axioms for propositional logic, such as $(P \vee P) \to P$, and certain rules of inference such as the classical rule of *modus ponens*: from P and $P \to Q$, infer Q. Post proved that all valid formulas of propositional logic follow from the axioms by means of these rules, so *Principia* is *complete* as far as propositional logic is concerned.

Post also showed that propositional logic is consistent, by introducing the now familiar device of truth tables. Truth tables assign to each axiom the value "true," and each rule of inference preserves the value "true," so all theorems have the value "true" and hence are true in the intuitive sense. The same idea also shows that propositional logic is consistent in the formal sense. That is, it does not prove any proposition P together with its negation $\neg P$, since if one of these has the value "true" the other has the value "false." Together, the two results solve what Post called the *finiteness problem* for propositional logic: to give an algorithm that determines, for any given proposition, whether it is a theorem.

We now know that propositional logic is far easier than the full *Principia*. Indeed Post's results were already known to Bernays and Hilbert in 1918, though not published (see, for example, Zach [**16**]). However, what is interesting is that Post went straight ahead, attempting to analyze *arbitrary* rules of inference. He took a "rule of inference" to consist of a finite set of *premises*

$$g_{11} P_{i_{11}} g_{12} P_{i_{12}} \ldots g_{1m_1} P_{i_{1m_1}} g_{1(m_1+1)}$$

$$g_{21} P_{i_{21}} g_{22} P_{i_{22}} \ldots g_{2m_2} P_{i_{2m_2}} g_{2(m_2+1)}$$

$$\ldots$$

$$g_{k1} P_{i_{k1}} g_{k2} P_{i_{k2}} \ldots g_{km_k} P_{i_{km_k}} g_{k(m_k+1)},$$

which together produce a *conclusion*

$$g_1 P_{i_1} g_2 P_{i_2} \ldots g_m P_{i_m} g_{m+1}.$$

The g_{ij} are certain specific symbols or strings of symbols, such as the $\rightarrow$ symbol in modus ponens, and the P_{kl} are arbitrary strings (such as the P and Q in modus ponens). Each P_{kl} in the conclusion is in at least one of the premises. Such "rules" include all the rules of *Principia* and, Post thought, any other rules that draw conclusions from premises in a determinate way.

The problem of analyzing such "production systems" amounts to understanding all possible formal systems, a task of seemingly overwhelming proportions. However, Post initially made surprising progress. By the end of the 1920–21 academic year he had proved his *normal form theorem*, which says that the theorems of any production system can be produced by a *normal system* with a single axiom and rules of only the simple form

$$gP \quad \text{produces} \quad Pg'.$$

In other words, any string beginning with the specific string g may be replaced by the string in which g is removed and g' is attached at the other end.

Normal systems include an even simpler class of systems that Post called "tag" systems, in which each g' depends only on the initial letter of g and all g have the same length. One such system uses only the letters 0 and 1, and each g has length 3. If g begins with 0, then $g' = 00$, and if g begins with 1 then $g' = 1101$. The left-hand end of the string therefore advances by three places at each step, trying to "tag" the right-hand end which advances by two or four places. For example, here is what happens when the initial string is 1010:

```
1010
   01101
      0100
         000
            00
               0
```

and then the string becomes empty. In all cases that Post tried, the result was either termination (as here) or periodicity, but he was unable to decide whether this was always the case. In fact, as far as I know the general behavior of this tag system is still not known. Post tried reducing the length of g and allowing more than two symbols, but it did not help.

> ... when this possibility was explored in the early summer of 1921, it rather led to an overwhelming confusion of classes of cases, with the solution of the corresponding problem depending more and more on problems in ordinary number theory. Since it had been our hope that the known difficulties of number theory would, as it were, be dissolved in the particularities of this more primitive form of mathematics, the solution of the general problem of "tag" appeared hopeless, and with it our entire program of the solution of finiteness problems. [**10**, p. 24]

After a few fruitless attempts to escape the difficulties with different normal forms, Post realized what the true situation must be: Theorems can indeed be produced by simple rules, but only because *any* computation can be reduced to simple steps. Predicting the outcome of simple rules, however, is no easier than deciding whether arbitrary sentences of mathematics are theorems. This fuller realization of the significance of the previous reductions led to a reversal of our entire program. [**10**, p. 44]

The reverse program was easier than the one he had set himself initially, which was essentially the following:

1. Describe all possible formal systems.
2. Simplify them.
3. Hence solve the deducibility problem for all of them.

Post's success in reducing complicated rules to simple ones convinced him that, for *any* system generating strings of symbols, there is a normal system that generates the same strings. But *it is possible to enumerate all normal systems*, since each consists of finitely many strings of symbols on a finite alphabet, and hence it is possible to enumerate all systems for generating theorems. This invites an application of the diagonal argument, described below. The outcome is that *for certain formal systems the deducibility problem is unsolvable*.

After this dramatic change of direction Post saw the true path as follows:

1. Describe all possible formal systems.
2. Diagonalize them.
3. Show that some of them have unsolvable deducibility problem.

And he also saw one step further—the incompleteness theorem—because:

4. No formal system obtains all the answers to an unsolvable problem.

Post's approach to incompleteness

We shall deal with Step 4 of Post's program first, because it is quite simple, and it dispels the myth that incompleteness is a difficult concept. Certainly, it rests on the concept of computability, but today we can define computability as "computable by a program in some standard programming language," and most readers will have a reasonable idea what this means.

Let us define an algorithmic problem, or simply *problem*, to be a computable list of questions:

$$P = \langle Q_1, Q_2, Q_3, \ldots \rangle$$

For example, the problem of recognizing primes is the list

$$\langle \text{"Is 1 prime?", "Is 2 prime?", "Is 3 prime?"}, \ldots \rangle$$

A problem is said to be *unsolvable* if the list of answers is not computable. The problem of recognizing primes is of course *solvable*.

Now suppose that an unsolvable $P = \langle Q_1, Q_2, Q_3, \ldots \rangle$ exists.

Then no consistent formal system F proves all correct sentences of the form

"The answer to Q_i is A_i.",

since by systematically listing all the theorems of F we could compute a list of answers to problem P.

Thus any consistent formal system F is *incomplete* with respect to sentences of the form "The answer to Q_i is A_i": there are some true sentences of this form that F does not prove.

It is true that there are several matters arising from this argument. What is the significance of consistency? Are there unprovable sentences in mainstream mathematics? But for Post incompleteness was a simple consequence of the existence of unsolvable problems. He also saw unsolvable problems as a simple consequence of the diagonal argument (described in the next section).

The really *big* problem, in Post's view, was to show that all computation is reflected in normal systems. Without a precise definition of computation, the concept of unsolvable problem is meaningless. Gödel was lucky not to be aware of this very general approach to incompleteness. His approach was to analyze *Principia Mathematica* (and "related systems") and prove its incompleteness directly. He did not see incompleteness as a consequence of unsolvability, in fact did not *believe* that computability could be precisely defined until he read Turing's paper [**14**], where the concept of Turing machine was defined.

Thus Post's proof of incompleteness was delayed because he was trying to do so much: The task he set himself in 1921 was in effect to do most of what Gödel, Church, and Turing did among them in 1931–36. In 1936, Church published a definition of computability [**1**] and gave the first published example of an unsolvable problem. But "Church's thesis"—that here was a precise definition of computability—was not accepted until the equivalent Turing machine concept appeared later in 1936, along with Turing's very lucid arguments for it.

As mentioned above, Post arrived at a similar concept independently [**9**], so in fact he completed his program in 1936. By then, unfortunately, it was too late for him to get credit for anything except a small share of the computability concept.

The diagonal argument

The diagonal argument is a very flexible way of showing the incompleteness of infinite lists: lists of real numbers, lists of sets of natural numbers, and lists of functions of natural numbers. It was perhaps implicit in Cantor's 1874 proof of the uncountability of the real numbers, but it first became clear and explicit in his 1891 proof, which goes as follows.

Suppose that $x_1, x_2, x_3, \ldots$ is a list of real numbers. More formally, suppose that to each natural number n there corresponds a real number x_n, and imagine a tabulation of the decimal expansions of these numbers one above the other, say

$$\begin{aligned} x_1 &= 3.\underline{1}4159\ldots \\ x_2 &= 2.7\underline{1}828\ldots \\ x_3 &= 1.41\underline{4}21\ldots \\ x_4 &= 0.577\underline{2}1\ldots \\ x_5 &= 1.6180\underline{3}\ldots \\ &\vdots \end{aligned}$$

A number x *not* on the list can always be constructed by making x differ from each x_n in the nth decimal place. For example, one can take the nth decimal place of x to be 1 if the nth decimal place of x_n is not 1, and 2 if the nth decimal place of x_n is 1. With the list above, we get the number

$$x = 0.22111\ldots.$$

The method for producing this new number x is called "diagonal," because it involves only the diagonal digits in the tabulation of $x_1, x_2, x_3, \ldots$.

It is commonly thought that the diagonal method is *nonconstructive*, but in fact the diagonal number x is clearly computable from the tabulation of $x_1, x_2, x_3, \ldots$. Indeed, one needs to compute only one decimal place of x_1, two decimal places of x_2, three decimal places of x_3, and so on. Turing observed that this tells us something interesting about computable real numbers [**14**].

It is *not* the case that there are uncountably many computable reals, because there are only countably many Turing machines (or programs in a fixed programming language, as we would prefer to define the concept of computation today) and at most one computable number is defined by each machine. Indeed, a real number is defined only if the machine behaves in a special way. In Turing's formulation the machine must print the successive digits of the number on specified squares of the machine's tape and must not change any digit once it is printed.

It would therefore seem, by the diagonal argument, that we could compute a number x *different* from each of the computable numbers $x_1, x_2, x_3, \ldots$. What is the catch?

There is no problem computing a list of all Turing machines, or programs. All of them are sequences of letters in a fixed finite alphabet, so they can be enumerated in lexicographic order. Also, once each machine is written down we can run it to produce digits of the number it defines, if any. The catch is that *we cannot identify all the machines that define computable real numbers*. The problem of recognizing all such machines is *unsolvable* in the sense that no Turing machine can correctly answer all the questions

Does machine 1 define a computable real?
Does machine 2 define a computable real?
Does machine 3 define a computable real? ....

There cannot be a Turing machine that solves this problem, otherwise we could hook it up to a machine that diagonalizes all the computable numbers and hence compute a number that is not computable.

What prevents the identification of machines that define computable numbers? When one explores this question, other unsolvable problems come to light. For example, we could try to catch all machines that fail to define real numbers by attaching to each one a device that halts computation as soon as the machine makes a misstep, such as changing a previously printed digit. As Turing pointed out, this implies the unsolvability of the *halting problem*: to decide, for any machine and any input, whether the machine eventually halts (or performs any other specific act). This problem is a perpetual thorn in the side of computer programmers, because it means that there is no general way to decide whether programs do what they are claimed to do. Unsolvable problems also arise in logic and mathematics, because systems such as predicate logic and number theory are capable of simulating all Turing machines. This is how Church and Turing proved the unsolvability of the *Entscheidungsproblem*, the problem of deciding validity of formulas in predicate logic.

Post's application of the diagonal argument Post also used the diagonal argument, but in the form used by Cantor (1891) to prove that any set has more subsets than elements. Given any set X, suppose each member $x \in X$ is paired with a subset $S_x \subseteq X$. Then the *diagonal* subset $D \subseteq X$ defined by

$$x \in D \leftrightarrow x \notin S_x$$

is different from each S_x, with respect to the element x.

The computable version of the diagonal argument takes X to be the set $\mathbb{N}$ of natural numbers, and for each $n \in \mathbb{N}$, what Post called the nth *recursively enumerable* subset S_n of $\mathbb{N}$. A recursively enumerable (r.e.) set is one whose members may be computably listed, and there are various ways to pair Turing machines with r.e. sets. For example, S_n may be defined as the set of input numbers m for which the nth machine has a halting computation. There is no loss of generality in considering the elements of an r.e. set to be numbers, because any string of symbols (in a fixed alphabet) can be encoded by a number.

A typical r.e. set is the set of theorems of a formal system, which is why Post was interested in the concept. Each theorem T is put into a machine, which systematically applies all rules of inference to the axioms, halting if and only if T is produced. Another example, which gives the flavor of the concept in a setting more familiar to mathematicians, consists of the strings of digits between successive 9s in the decimal expansion of π. Since

$$\pi = 3.14159\underline{265358}9\underline{7}9\underline{323846264338327}9\underline{5028841}9\underline{716}9\underline{3}9937510\ldots,$$

the set in question is

$$S = \{265358, 7, 323846264338327, 5028841, 716, 3, \ldots\}.$$

It is clear that a list of members of S can be computed, since π is a computable number, but otherwise S is quite mysterious. We do not know how to decide membership for S, or even whether S is infinite. This is typical of r.e. sets, and useful to keep in mind when constructing r.e. sets that involve arbitrary computations.

The diagonal set D is not r.e., being different from the nth r.e. set S_n with respect to the number n; however, its complement $\overline{D}$ *is* r.e. This is because

$$n \in \overline{D} \leftrightarrow n \in S_n,$$

so any $n \in \overline{D}$ will eventually be found by running the nth machine on input n. Thus *$\overline{D}$ is an example of an r.e. set whose complement is not r.e.*. It follows that no machine can decide, for each n, whether $n \in \overline{D}$ (or equivalently, whether $n \in S_n$). If there were such a machine, we could list all the members of D by asking

Is $1 \in S_1$?
Is $2 \in S_2$?
Is $3 \in S_3$?
...

and collecting the n for which the answer is no.

It also follows that *no consistent formal system can prove all theorems of the form $n \notin S_n$*, since this would yield a listing of D. This is a version of the incompleteness theorem, foreseen by Post in 1921, but first published by Gödel in 1931 [**4**].

Differences between Post and Gödel

As we have seen, Post's starting point was the concept of computation, which he believed could be formalized and made subject to the diagonal argument. Diagonalization yields problems that are *absolutely* unsolvable, in the sense that no computation can solve them. In turn, this leads to *relatively* undecidable propositions, for example, propositions of the form $n \notin S_n$. No consistent formal system F can prove all true propositions of this form, hence any such F must fail to prove some true proposition $n_0 \notin S_{n_0}$. But this proposition is only relatively undecidable, not absolutely, because F can be consistently extended by adding it as an axiom.

Gödel did not at first believe in absolutely unsolvable problems, because he did not believe that computation is a mathematical concept. Instead, he proved the existence of relatively undecidable propositions directly, by constructing a kind of diagonal argument inside *Principia Mathematica.* Also, he *arithmetized* the concept of proof there, so provability is expressed by a number-theoretic relation, and his undecidable proposition belongs to number theory. Admittedly, Gödel's proposition is not otherwise interesting to number theorists, but Gödel saw that it is interesting for another reason: *it expresses the consistency of Principia Mathematica.*

This remarkable fact emerges when one pinpoints the role, in the incompleteness proof, of the assumption that the formal system F is consistent, as we will soon explain. It seems that Gödel deserves full credit for this observation, which takes logic even higher than the level reached with the discovery of incompleteness.

Outsmarting a formal system We now reflect on Post's incompleteness proof for a formal system F, to find an explicit n_0 such that $n_0 \notin S_{n_0}$ is true but not provable by F.

It is necessary to assume that F is consistent, because an inconsistent formal system (with a modicum of ordinary logic) proves everything. In fact it is convenient to assume more, namely, that F proves only true propositions. Now consider the r.e. set of propositions of the form $n \notin S_n$ proved by F. The corresponding numbers n also form an r.e. set, with index n_0 say. That is,

$$S_{n_0} = \{n : F \text{ proves } n \notin S_n\}.$$

By definition of S_{n_0}, $n_0 \in S_{n_0}$ implies that F proves the proposition $n_0 \notin S_{n_0}$. But if so, $n_0 \notin S_{n_0}$ is true, and we have a contradiction. Thus the truth is that $n_0 \notin S_{n_0}$, *but F does not prove this fact.*

It seems that we know more than F, but how come? The "extra smarts" needed to do better than F lie in the ability to recognize that F is consistent (or, strictly speaking, that all theorems of F are true). In fact, what we have actually proved is the theorem

$$\text{Con}(F) \rightarrow n_0 \notin S_0,$$

where $\text{Con}(F)$ is a proposition that expresses the consistency of F. It follows that $\text{Con}(F)$ *is not provable in* F, otherwise the proposition $n_0 \notin S_{n_0}$ would also be provable (by modus ponens). But if we can "see" $\text{Con}(F)$, then we can "see" $n_0 \notin S_{n_0}$.

If F is a really vast system, like *Principia Mathematica* or a modern system of set theory, then it takes a lot of chutzpah to claim the ability to see $\text{Con}(F)$. But the incompleteness argument also applies to modest systems of number theory, which everybody believes to be consistent, *because we know an interpretation of the axioms*: $1, 2, 3, \ldots$ stand for the natural numbers, $+$ stands for addition, and so on. Thus the ability to see meaning in a formal system F actually confers an advantage: It allows us to see $\text{Con}(F)$, and hence to see propositions not provable by F.

Now recall how this whole story began. *Principia Mathematica* and other formal systems F were constructed in the belief that there was everything to gain (in rigor, precision, and clarity) and nothing to lose in treating deduction as computation with meaningless symbols. Gödel showed that this is not the case. Loss of meaning causes loss of theorems, such as $\text{Con}(F)$. It is surprising how little this is appreciated. More than 60 years ago Post wrote:

> It is to the writer's continuing amazement that ten years after Gödel's remarkable achievement current views on the nature of mathematics are thereby affected only to the point of seeing the need of many formal systems, instead of a universal one. Rather has it seemed to us to be inevitable that these developments will

> result in a reversal of the entire axiomatic trend of the late 19th and early 20th centuries, with a return to meaning and truth. [**10**, p. 378]

Perhaps it is too much to expect a "reversal of the entire axiomatic trend," but a milder proposal seems long overdue. Post's words should be remembered every time we plead with our students not to manipulate symbols blindly, but to understand what they are doing.

Acknowledgments. The biographical information in this article is drawn mainly from Martin Davis's introduction to Post's collected works [**2**] and the web site of the American Philosophical Society, where most of Post's papers are held. I am also indebted to Martin Davis and John Dawson for supplying me with photocopies of correspondence between Post and Gödel.

REFERENCES

1. A. Church, An unsolvable problem in elementary number theory, *Amer. J. Math.* **58** (1936), 345–363.
2. M. Davis, ed., *Solvability, Provability, Definability: The Collected Works of Emil L. Post.*, Birkhäuser, 1994.
3. ———, *The Universal Computer*, W. W. Norton and Company, 2000.
4. K. Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik* **38** (1931), 173–198.
5. G. Higman, Subgroups of finitely presented groups, *Proc. Roy. Soc. London Ser. A* **262** (1961), 455–475.
6. A. Markov, Unsolvability of the problem of homeomorphy (Russian), *Doklady Akademiĭ Nauk SSSR* **121** (1958), 218–220.
7. P. S. Novikov, On the algorithmic unsolvability of the word problem in group theory (Russian), *Doklady Akademiĭ Nauk SSSR* Mat. Inst. Tr. **44** (1955).
8. E. L. Post, Introduction to the general theory of elementary propositions, *Amer. J. Math.* **43** (1921), 163–185.
9. ———, Finite combinatory processes, Formulation 1., *J. Symbolic Logic* **1** (1936), 103–105.
10. ———, Absolutely unsolvable problems and relatively undecidable propositions—account of an anticipation, in Davis [**2**], 375–441.
11. ———, Recursively enumerable sets of positive integers and their decision problems, *Bull. Amer. Math. Soc.* **50** (1944), 284–316.
12. ———, A variant of a recursively unsolvable problem, *Bull. Amer. Math. Soc.* **52** (1945), 264–268.
13. ———, Recursive unsolvability of a problem of Thue, *J. Symbolic Logic* **12** (1947), 1–11.
14. A. M. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proc. of the London Math. Soc.* **42** (1936), 230–265.
15. A. N. Whitehead and B. Russell, *Principia Mathematica*, Cambridge University Press, 3 vols. 1910–1913.
16. R. Zach, *Hilbert's Finitism: Historical, Philosophical, and Metamathematical Perspectives*, UC Berkeley Ph.D. thesis in Logic and Methodology of Science 2001, on the web at `http://www.ucalgary.ca/~rzach/papers/hilbert.html`

Enter, Stage Center: The Early Drama of the Hyperbolic Functions

JANET HEINE BARNETT
University of Southern Colorado
(Colorado State University—Pueblo)
Pueblo, CO 81001–4901
janet.barnett@colostate-pueblo.edu

In addition to the standard definitions of the hyperbolic functions (for instance, $\cosh x = (e^x + e^{-x})/2$), current calculus textbooks typically share two common features: a comment on the applicability of these functions to certain physical problems (for instance, the shape of a hanging cable known as the catenary) and a remark on the analogies that exist between properties of the hyperbolic functions and those of the trigonometric functions (for instance, the identities $\cosh^2 x - \sinh^2 x = 1$ and $\cos^2 x + \sin^2 x = 1$). Texts that offer historical sidebars are likely to credit development of the hyperbolic functions to the 18th-century mathematician Johann Lambert. Implicit in this treatment is the suggestion that Lambert and others were interested in the hyperbolic functions in order to solve problems such as predicting the shape of the catenary. Left hanging is the question of whether hyperbolic functions were developed in a deliberate effort to find functions with trig-like properties that were required by physical problems, or whether these trig-like properties were unintended and unforeseen by-products of the solutions to these physical problems. The drama of the early years of the hyperbolic functions is far richer than either of these plot lines would suggest.

Prologue: The catenary curve

What shape is assumed by a flexible inextensible cord hung freely from two fixed points? Those with an interest in the history of mathematics would guess (correctly) that this problem was first resolved in the late 17th century and involved the Bernoulli family in some way. The curve itself was first referred to as the "catenary" by Huygens in a 1690 letter to Leibniz, but was studied as early as the 15th century by da Vinci. Galileo mistakenly believed the curve would be a parabola [**8**]. In 1669, the German mathematician Joachim Jungius (1587–1657) disproved Galileo's claim, although his correction does not seem to have been widely known within 17th-century mathematical circles.

17th-century mathematicians focused their attention on the problem of the catenary when Jakob Bernoulli posed it as a challenge in a 1690 *Acta Eruditorum* paper in which he solved the isochrone problem of constructing the curve along which a body will fall in the same amount of time from any starting position. Issued at a time when the rivalry between Jakob and Johann Bernoulli was still friendly, this was one of the earliest challenge problems of the period. In June 1691, three independent solutions appeared in *Acta Eruditorum* [**1, 11, 16**]. The proof given by Christian Huygens employed geometrical arguments, while those offered by Gottfried Leibniz and Johann Bernoulli used the new differential calculus techniques of the day. In modern terminology, the crux of Bernoulli's proof was to show that the curve in question satisfies the differential equation $dy/dx = s/k$, where s represents the arc length from the vertex P to an arbitrary point Q on the curve and k is a constant depending on the weight per unit length of cord as in FIGURE 1.

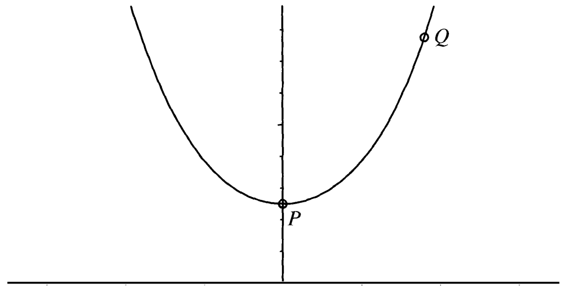


Figure 1 The catenary curve

Showing that $y = k\cosh(x/k)$ is a solution of this differential equation is an accessible problem for today's second-semester calculus student. 17th-century solutions of the problem differed from those of today's calculus students in a particularly notable way: *There was absolutely no mention of hyperbolic functions, or any other explicit function, in the solutions of 1691!* In these early days of calculus, curve constructions, and not explicit functions, were cast in the leading roles.

A suggestion of this earlier perspective can be heard in a letter dated September 19, 1718 sent by Johann Bernoulli to Pierre Réymond de Montmort (1678–1719):

> The efforts of my brother were without success; for my part, I was more fortunate, for I found the skill (I say it without boasting, why should I conceal the truth?) to solve it in full and to reduce it to the rectification of the parabola. It is true that it cost me study that robbed me of rest for an entire night. It was much for those days and for the slight age and practice I then had, but the next morning, filled with joy, I ran to my brother, who was still struggling miserably with this Gordian knot without getting anywhere, always thinking like Galileo that the catenary was a parabola. Stop! Stop! I say to him, don't torture yourself any more to try to prove the identity of the catenary with the parabola, since it is entirely false. The parabola indeed serves in the construction of the catenary, but the two curves are so different that one is algebraic, the other is transcendental ... (as quoted by Kline [**13**, p. 473]).

The term *rectification* in this passage refers to the problem of determining the arc length of a curve. The particular parabola used in Bernoulli's construction (given by $y = x^2/8 + 1$ in modern notation) was defined geometrically by Bernoulli as having "latus rectum quadruple the latus rectum of an equilateral hyperbola that shares the same vertex and axis" [**1**, pp. 274–275]. Bernoulli used the arc length of the segment of this parabola between the vertex $B = (0, 1)$ and the point $H = (\sqrt{8(y-1)}, y)$ to construct a segment GE such that the point E would lie on the catenary. In modern notation, the length of segment GE is the parabolic arc length BH, given by

$$\text{Arclength} = \sqrt{y^2-1} + \ln\left(y + \sqrt{y^2-1}\right),$$

while the catenary point E is given by

$$E = \left(-\ln\left(y+\sqrt{y^2-1}\right), y\right) = \left(x, \frac{e^x+e^{-x}}{2}\right).$$

The expression $\sqrt{y^2-1}$ in the arc length formula is the abscissa of the point $G(\sqrt{y^2-1}, y)$ on the equilateral hyperbola ($y^2 - x^2 = 1$) that played both the central role described above in defining the parabola necessary for the construction, as well as a supporting role in constructing the point E. Because a procedure for rectifying a parabola was known by this time, this reduction of the catenary problem to the rectification of a parabola provided a complete 17th-century solution to the catenary problem.

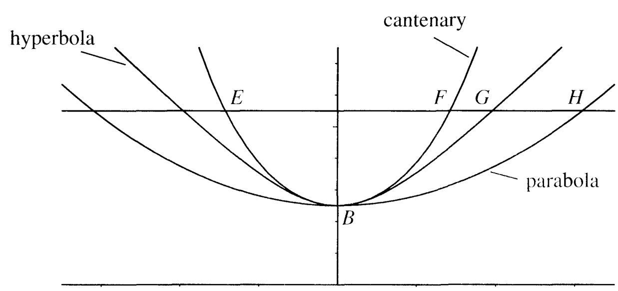


Figure 2 Bernoulli's construction of the catenary curve

Interestingly, another of the "first solvers" of the catenary problem, Christian Huygens, solved the rectification problem for the parabola as early as 1659. In fact, although the rectification problem had been declared by Descartes as beyond the capacity of the human mind [**4**, pp. 90–91], the problem of rectifying a curve C was known to be equivalent to the problem of finding the area under an associated curve C' by the time Huygens took up the parabola question.

A general procedure for determining the curve C' was provided by Hendrick van Heuraet (1634–1660) in a paper that appeared in van Schooten's 1659 Latin edition of Descartes' *La Geometrie*. (In modern notation, C' is defined by $L(t) = \int_a^t \sqrt{1 + (dy/dx)^2}\,dx$, where $y = f(x)$ defines the original curve C.) Huygens used this procedure to show that rectification of a parabola is equivalent to finding the area under a hyperbola. A solution of this latter problem in the study of curves—determining the area under a hyperbola—was first published by Gregory of St. Vincent in 1647 [**13**, p. 354]. Anton de Sarasa later recognized (in 1649) that St. Vincent's solution to this problem provided a method for computation of logarithmic values.

As impressive as these early "pre-calculus" calculus results were, by the time the catenary challenge was posed by Jakob Bernoulli in 1690, the rate at which the study of curves was advancing was truly astounding, thanks to the groundbreaking techniques that had since been developed by Isaac Newton (1642–1727) and Gottfried Leibniz (1646–1716). Relations between the Bernoulli brothers fared less well over the ensuing decades, as indicated by a later passage from Johann Bernoulli's 1718 letter to Montmort:

> But then you astonish me by concluding that my brother found a method of solving this problem. ... I ask you, do you really think, if my brother had solved the problem in question, he would have been so obliging to me as not to appear among the solvers, just so as to cede me the glory of appearing alone on the stage in the quality of the first solver, along with Messrs. Huygens and Leibniz? (as quoted by Kline [**13**, p. 473])

Historical evidence supports Johann's claim that Jakob was not a "first solver" of the catenary problem. But in the year immediately following that first solution, Jakob Bernoulli and others solved several variations of this problem. Huygens, for example, used physical arguments to show that the curve is a parabola if the total load of cord and suspended weights is uniform per horizontal foot, while for the true catenary, the weight per foot along the cable is uniform. Both Bernoulli brothers worked on determining the shape assumed by a hanging cord of variable density, a hanging cord of constant thickness, and a hanging cord acted on at each point by a force directed to a fixed center. Johann Bernoulli also solved the converse problem: given the shape assumed by a flexible inelastic hanging cord, find the law of variation of density of the cord. Another nice result due to Jakob Bernoulli stated that, of all shapes that may be assumed by flexible inelastic hanging cord, the catenary has the lowest center of gravity.

A somewhat later appearance of the catenary curve was due to Leonhard Euler in his work on the calculus of variations. In his 1744 *Methodus Inveniendi Lineas Curvas Maximi Minimive Proprietate Gaudentes* [**5**], Euler showed that a catenary revolved about its axis (the catenoid) generates the only minimal surface of revolution. Calculating the surface area of this minimal surface is another straightforward exercise that can provide a nice historical introduction to the calculus of variations for second-semester calculus students. Kline [**13**, p. 579] comments that Euler himself did not make effective use of the full power of the calculus in the *Methodus*; derivatives were replaced by difference quotients, integrals by finite sums, and extensive use was made of geometric arguments. In tracing the story of the hyperbolic functions, this last point cannot be emphasized enough. From its earliest introduction in the 15th century through Euler's 1744 result on the catenoid, there is no connection made between analytic expressions involving the exponential function and the catenary curve. Indeed, prior to the development of 18th-century analytic techniques, no such connection could have been made. Calculus in the age of the Bernoullis was "the Calculus of Curves," and the catenary curve is just that—a *curve*. The hyperbolic functions did not, and could not, come into being until the full power of formal analysis had taken hold in the age of Euler.

Act I: The hyperbolic functions in Euler?

In seeking the first appearance of the hyperbolic functions as *functions*, one naturally looks to the works of Euler. In fact, the expressions $(e^x + e^{-x})/2$ and $(e^x - e^{-x})/2$ do make an appearance in Volume I of Euler's *Introductio in analysin infinitorum* (1745, 1748) [**6**]. Euler's interest in these expressions seems natural in view of the equations $\cos x = (e^{\sqrt{-1}x} + e^{-\sqrt{-1}x})/2$ and $\sqrt{-1}\sin x = (e^{\sqrt{-1}x} - e^{-\sqrt{-1}x})/2$ that he derived in this text. However, Euler's interest in what we call hyperbolic functions appears to have been limited to their role in deriving infinite product representations for the sine and cosine functions. Euler did not use the word *hyperbolic* in reference to the expressions $(e^x + e^{-x})/2$, $(e^x - e^{-x})/2$, nor did he provide any special notation or name for them. Nevertheless, his use of these expressions is a classic example of Eulerian analysis, included here as an illustration of 18th-century mathematics. An analysis of this derivation, either in its historical form or in modern translation, would be suitable for student projects in pre-calculus and calculus, or as part of a mathematics history course.

To better illustrate the style of Euler's analysis and the role played within it by the hyperbolic expressions, we employ his notation from the *Introductio* throughout this section. Although sufficiently like our own to make the work accessible to modern readers, there are interesting differences. For instance, Euler's use of periods in "sin . x" and "cos . x" suggests the notation still served as abbreviations for *sinus* and *cosinus*, rather than as symbolic function names. Like us, Euler and his contemporaries were intimately familiar with the infinite series representations for sin . x and cos . x, but generally employed infinite series with less than the modern regard for rigor. Thus, as established in Section 123 of the *Introductio*, Euler could (and did) rewrite the expression $(e^x - e^{-x})$ as

$$e^x - e^{-x} = \left(1 + \frac{x}{i}\right)^i - \left(1 - \frac{x}{i}\right)^i = 2\left(\frac{x}{1} + \frac{x^3}{1\cdot 2\cdot 3} + \frac{x^5}{1\cdot 2\cdot 3\cdot 4\cdot 5} + \text{etc.}\right),$$

where i represented an infinitely large quantity (and *not* the square root of -1, denoted throughout the *Introductio* as $\sqrt{-1}$). Other results used by Euler are also familiarly unfamiliar to us, most notably the fact that $a^n - z^n$ has factors of the form $aa - 2az\cos .2k\pi/n + zz$, as established in Section 151 of the *Introductio.*

Euler's development of infinite product representations for $\sin . x$ and $\cos . x$ in the *Introductio* begins in Section 156 by setting $n = i$, $a = 1 + x/i$, and $z = 1 - x/i$ in the expression $a^n - z^n$ (where, again, i is infinite, so, for example, $a^n = (1 + x/i)^i = e^x$). After some algebra, the result of Section 151 cited above allowed Euler to conclude that $e^x - e^{-x}$ has factors of the form $2 - 2xx/ii - 2(1 - xx/ii)\cos .2k\pi/n$. Substituting $\cos .2k\pi/n = 1 - (2kk/ii)\pi\pi$ (the first two terms of the infinite series representation for cosine) into this latter expression and doing a bit more algebra, Euler obtained the equation

$$2 + \frac{2xx}{ii} - 2\left(1 - \frac{xx}{ii}\right)\cos .2k\frac{\pi}{n} = \frac{4xx}{ii} + \left(\frac{4kk}{ii}\right)\pi\pi - \frac{4kk\pi\pi xx}{i^4}.$$

Ergo (to quote Euler), $e^x - e^{-x}$ has factors of the form $1 + xx/(kk\pi\pi) - xx/ii$. Since i is an infinitely large quantity, Euler's arithmetic of infinite and infinitesimal numbers allowed the last term to drop out. (Tuckey and McKenzie give a thorough discussion of these ideas [**17**].) The end result of these calculations, as presented in Section 156, thereby became

$$\frac{e^x - e^{-x}}{2} = x\left(1 + \frac{xx}{\pi\pi}\right)\left(1 + \frac{xx}{4\pi\pi}\right)\left(1 + \frac{xx}{9\pi\pi}\right)\left(1 + \frac{xx}{16\pi\pi}\right)\text{etc.}$$
$$= 1 + \frac{xx}{1\cdot 2\cdot 3} + \frac{x^4}{1\cdot 2\cdot 3\cdot 4\cdot 5} + \frac{x^6}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7} + \text{etc.} \quad (1)$$

A similar calculation (Section 157) derived the analogous series for $(e^x + e^{-x})/2$.

In Section 158, Euler employed these latter two results in the following manner. Recalling the well-known fact (which he derived in Section 134) that

$$\frac{e^{z\sqrt{-1}} - e^{z\sqrt{-1}}}{2\sqrt{-1}} = \sin . z = z - \frac{z^3}{1\cdot 2\cdot 3} + \frac{z^5}{1\cdot 2\cdot 3\cdot 4\cdot 5} - \text{etc.},$$

Euler let $x = z\sqrt{-1}$ in equation (1) above to get

$$\sin .z = z\left(1 - \frac{zz}{\pi\pi}\right)\left(1 - \frac{zz}{4\pi\pi}\right)\left(1 - \frac{zz}{9\pi\pi}\right)\left(1 - \frac{zz}{16\pi\pi}\right)\text{etc.}$$
$$= z\left(1 - \frac{z}{\pi}\right)\left(1 + \frac{z}{\pi}\right)\left(1 - \frac{z}{2\pi}\right)\left(1 - + \frac{z}{2\pi}\right)\left(1 - \frac{z}{3\pi}\right)\text{etc.}$$

The same substitution, applied to the series with even terms, yielded the now-familiar product representation for $\cos .z$.

Here we arrive at Euler's apparent goal: the derivation of these lovely infinite product representations for the sine and cosine. Although the expressions $(e^x + e^{-x})/2$ and

$(e^x - e^{-x})/2$ played a role in obtaining these results, it was a supporting role, with the arrival of the hyperbolic functions on center stage yet to come.

Act II, Scene I: Lambert's first introduction of hyperbolic functions

Best remembered today for his proof of the irrationality of π, and considered a forerunner in the development of noneuclidean geometries, Johann Heinrich Lambert was born in Mülhasen, Alsace on August 26, 1728. The Lambert family had moved to Mülhasen from Lorraine as Calvinist refugees in 1635. His father and grandfather were both tailors. Because of the family's impoverished circumstances (he was one of seven children), Lambert left school at age 12 to assist the family financially. Working first in his father's tailor shop and later as a clerk and private secretary, Lambert accepted a post as a private tutor in 1748 in the home of Reichsgraf Peter von Salis. As such, he gained access to a good library that he used for self-improvement until he resigned his post in 1759. Lambert led a largely peripatetic life over the next five years. He was first proposed as a member of the Prussian Academy of Sciences in Berlin in 1761. In January 1764, he was welcomed by the Swiss community of scholars, including Euler, in residence in Berlin. According to Scriba [**21**], Lambert's appointment to the Academy was delayed due to "his strange appearance and behavior." Eventually, he received the patronage of Frederick the Great (who at first described him as "the greatest blockhead") and obtained a salaried position as a member of the physical sciences section of the Academy on January 10, 1765. He remained in this position, regularly presenting papers to each of its divisions, until his death in 1777 at the age of 49.

Lambert was a prolific writer, presenting over 150 papers to the Berlin Academy in addition to other published and unpublished books and papers written in German, French, and Latin. These included works on philosophy, logic, semantics, instrument design, land surveying, and cartography, as well as mathematics, physics, and astronomy. His interests appeared at times to shift almost randomly from one topic to another, and often fell outside the mainstream of 18th-century science and mathematics. We leave it to the reader to decide whether his development of the hyperbolic functions is a case in point, or an exception to this tendency.

Lambert first treated hyperbolic trigonometric functions in a paper presented to the Berlin Academy of Science in 1761 that quickly became famous: *Mémoire sur quelques propriétés remarquables des quantités transcendantes circulaires et logarithmiqes* [**14**]. Rather than its consideration of hyperbolic functions, this paper was (and is) celebrated for giving the first proof of the irrationality of π. Lambert established this long-awaited result using continued fractions representations to show that z and $\tan z$ cannot both be rational; thus, since $\tan(\pi/4)$ is rational, π can not be.

Instead of concluding the paper at this rather climatic point, Lambert turned his attention in the last third of the paper to a comparison of the "*transcendantes circulaire*" [$\sin v$, $\cos v$,] with their analogues, the "*quantités transcendantes logarithmiques*" [$(e^v + e^{-v})/2$, $(e^v - e^{-v})/2$]. Beginning in Section 73, he first noted that the transcendental logarithmic quantities can be obtained from the transcendental circular quantities by taking all the signs in

$$\sin v = v - \frac{1}{2\cdot 3}v^3 + \frac{1}{2\cdot 3\cdot 4\cdot 5}v^5 - \frac{1}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7}v^7 + \text{etc.}$$

to be positive, thereby obtaining

$$\frac{e^v - e^{-v}}{2} = v + \frac{1}{2\cdot 3}v^3 + \frac{1}{2\cdot 3\cdot 4\cdot 5}v^5 + \frac{1}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7}v^7 + \text{etc.},$$

and similarly for the cosine series. He then derived continued fraction representations (in Section 74) for the expressions $(e^v - e^{-v})/2$, $(e^v + e^{-v})/2$, and $(e^v - e^{-v})/(e^v + e^{-v})$, and noted that these continued fraction representations can be used to show that v and e^v cannot both be rational. The fact that none of its powers or roots are rational prompted Lambert to speculate that e satisfied *no* algebraic equation with rational coefficients, and hence is *transcendental*. Charles Hermite (1822–1901) finally proved this fact in 1873. (Ferdinand Lindemann (1852–1939) established the transcendence of π in 1882.)

Although Lambert did not introduce special notation for his "*quantités transcendantes logarithmiques*" in this paper, he did go on to develop the analogy between these functions and the circular trigonometric functions that he said "should exist" because

> ... the expressions $e^u + e^{-u}$, $e^u - e^{-u}$, by substituting $u = v\sqrt{-1}$, give the circular quantities $e^{v\sqrt{-1}} + e^{-v\sqrt{-1}} = 2\cos v$, $e^{v\sqrt{-1}} - e^{-v\sqrt{-1}} = 2\sin v \cdot \sqrt{-1}$.

Lambert was especially interested in developing this "affinity" as far as possible without introducing imaginary quantities. To do this he introduced (in Section 75) a parameterization of an "equilateral hyperbola" ($x^2 - y^2 = 1$) to define the hyperbolic functions in a manner directly analogous to the definition of trigonometric functions by means of a unit circle ($x^2 + y^2 = 1$). Lambert's parameter is twice the area of the hyperbolic sector shown in FIGURE 3. Lambert used the letter M to denote a typical point on the hyperbola, with coordinates (ξ, η).

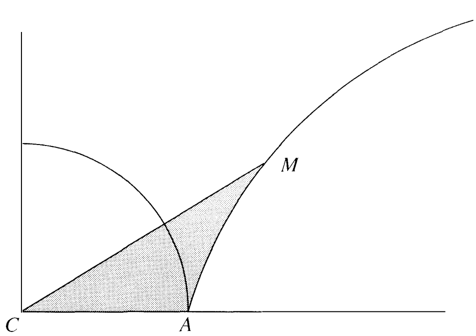


Figure 3 The parameter u represents twice the area of the shaded sector MCA

In Lambert's own diagram (FIGURE 4), the circle and the hyperbola are drawn together. The letter C marks the common center of the circle and the hyperbola, CA is the radius of the circle, CF the asymptote of the hyperbola, and AB the tangent line common to the circle and the hyperbola. The typical point on the hyperbola corresponds to a point N on the circle, with coordinates (x, y). Lowercase letters m and n mark nearby points on the hyperbola and circle, for use in differential computations.

Denoting the angle MCA by ϕ, Lambert listed several differential properties for quantities defined within this diagram, using a two-columned table intended to display the similarities between the "*logarithmiques*" and "*circulaires*" functions. The first seven lines of this table, reproduced below, defined the necessary variables and stated basic algebraic and trigonometric relations between them. Note especially the third line of this table, where $u/2$ (which Lambert denoted as $u : 2$) is defined to be the area of the hyperbolic "segment" $AMCA$.

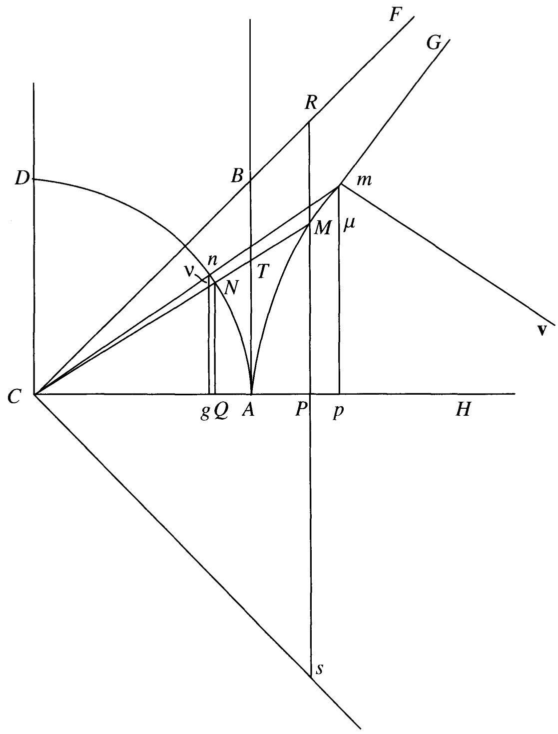


Figure 4 Diagram from Lambert's 1761 *Mémoire*

	pour l'hyperbole	*pour le cercle*
l'abscisse	$CP = \xi \ldots$	$\ldots CQ = x$
l'ordonné	$PM = \eta \ldots$	$\ldots QN = y$
le segment	$AMCA = u : 2 \ldots$	$\ldots ANCA = v : 2$
et il sera		
	$\text{tang}\,\phi = \frac{\eta}{\xi} \ldots$	$\ldots \text{tang}\,\phi = \frac{y}{x}$
	$1 + \eta\eta = \xi\xi = \eta\eta \cot\phi^2 \ldots$	$\ldots 1 - yy = xx = yy \cot\phi^2$
	$\xi\xi - 1 = \eta\eta = \xi\xi\, \text{tang}\,\phi^2 \ldots$	$\ldots 1 - xx = yy = xx\, \text{tang}\,\phi^2$
	$CM^2 = \xi^2 + \eta^2$	$CN^2 = x^2 + y^2$
	$= \xi^2(1 + \text{tang}\,\phi^2) = \frac{1+\text{tang}\,\phi^2}{1-\text{tang}\,\phi^2}$	$= x^2(1 + \text{tang}\,\phi^2) = \frac{1+\text{tang}\,\phi^2}{1+\text{tang}\,\phi^2} = 1$

Using these relations, it is a straightforward exercise to derive expressions for the differentials $d\xi$, $d\eta$, dx, and dy (as a step toward finding infinite series expressions for ξ and η). For example, given $\xi\xi - 1 = \eta\eta = \xi\xi\, \text{tang}\,\phi^2$ (tang would be tan in modern notation), it follows that $\xi = 1/\sqrt{1 - \text{tang}\,\phi^2}$. Lambert noted this fact, along with the differential $d\xi = \text{tang}\,\phi\, d\,\text{tang}\,\phi/(1 - \text{tang}\,\phi^2)^{3:2}$ obtained from it, later in the table.

To see how differential expressions for du and dv might be obtained, note that u is defined to be twice the area of the hyperbolic sector *AMCA*. The differential du thus represents twice the area of the hyperbolic sector *MCm*. This differential sector can be approximated by the area of a circular sector of radius *CM* and angle $d\phi$; that is, $du = 2[CM^2 d\phi/2]$. Substituting $CM^2 = (1 + \text{tang}\,\phi^2)/(1 - \text{tang}\,\phi^2)$ from the table above then yields $du = d\phi \cdot CM^2 = d\phi \cdot (1 + \text{tang}\,\phi^2)/(1 - \text{tang}\,\phi^2)$, where $d\phi \cdot (1 + \text{tang}\,\phi^2) = d(\text{tang}\,\phi)$. Thus, $du = d\,\text{tang}\,\phi/(1 - \text{tang}\,\phi^2)$. Although Lambert omitted the details of these derivations, his table summarized them as shown below.

pour l'hyperbole	*pour le cercle*
$+\,du = d\phi \cdot \left(\frac{1+\text{tang}\,\phi^2}{1-\text{tang}\,\phi^2}\right)$	$dv = d\phi = \frac{d\,\text{tang}\,\phi}{1+\text{tang}\,\phi^2}$
$= \frac{d\,\text{tang}\,\phi}{1-\text{tang}\,\phi^2}$	
$+\,d\xi = \frac{\text{tang}\,\phi\, d\,\text{tang}\,\phi}{(1-\text{tang}\,\phi^2)^{3:2}}$	$-\,dx = \frac{\text{tang}\,\phi\, d\,\text{tang}\,\phi}{(1+\text{tang}\,\phi^2)^{3:2}}$
$+\,d\eta = \frac{d\,\text{tang}\,\phi}{(1-\text{tang}\,\phi^2)^{3:2}}$	$+\,dy = \frac{d\,\text{tang}\,\phi}{(1+\text{tang}\,\phi^2)^{3:2}}$
$\vdots$	$\vdots$
$+\,d\xi : du = \eta \ldots$	$\ldots - dx : dv = y$
$+\,d\eta : du = \xi \ldots$	$\ldots + dy : dv = x$
$+\,d\xi = d\eta \cdot \text{tang}\,\phi \ldots$	$\ldots - dx : dy = \text{tang}\,\phi$

Using the relations $+\,d\xi : du = \eta$, $+\,d\eta : du = \xi$ from this table, along with standard techniques of the era for determining the coefficients of infinite series, Lambert then proved (Section 77) that the following relations hold:

$$\eta = u + \frac{1}{2\cdot 3}u^3 + \frac{1}{2\cdot 3\cdot 4\cdot 5}u^5 + \frac{1}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7}u^7 + \text{etc.}$$

$$\xi = u + \frac{1}{2}u^2 + \frac{1}{2\cdot 3\cdot 4}u^4 + \frac{1}{2\cdot 3\cdot 4\cdot 5\cdot 6}u^6 + \text{etc.},$$

where we recall that ξ is the abscissa of a point on the hyperbola, η is the ordinate of that same point, and u represents twice the area of the hyperbolic segment determined by that point. But these are exactly the infinite series for $(e^u - e^{-u})/2$ and $(e^u + e^{-u})/2$ with which Lambert began his discussion of the "*quantités transcendantes logarithmiques.*"

Lambert was thus able to conclude (Section 78) that $\xi = (e^u - e^u)/2$ and $\eta = (e^u - e^{-u})/2$ are, respectively, the abscissa and ordinate of a point on the hyperbola for which u represents twice the area of the hyperbolic segment determined by that point.

A derivation of this result employing integration, as outlined in some modern calculus texts, is another nice problem for students. Contrary to the suggestion of some texts, it is this parameterization of the hyperbola by the hyperbolic sine and cosine, and the analogous parameterization of the circle by the circular sine and cosine, that seems to have motivated Lambert and others eventually to provide the hyperbolic functions with trig-like names—not the similarity of their analytic identities. This is not to say that the similarities between the circular identities and the hyperbolic identities were without merit in Lambert's eyes—we shall see that Lambert and others exploited these similarities for various purposes. But Lambert's immediate interest in his 1761 paper lay elsewhere, as we shall examine more closely in the following section.

Interlude: Giving credit where credit is due

As Lambert himself remarked at several points in his 1761 *Mémoire*, he was especially interested in developing the analogy between the two classes of functions (circular versus hyperbolic) as far as possible without the use of imaginary quantities, and it is the geometric representation (that is, the parameterization) that provides him a means to this end. Lambert ascribed his own interest in this theme to the work of another 18th-century mathematician whose name is less well known, Monsieur le Chevalier François Daviet de Foncenex.

As a student at the Royal Artillery School of Turin, de Foncenex studied mathematics under a young Lagrange. As recounted by Delambre, the friendships Lagrange formed with de Foncenex and other students led to the formation of the Royal Academy of Science of Turin [**7**]. A major goal of the society was the publication of mathematical and scientific papers in their *Miscellanea Taurinensia*, or *Mélanges de Turin*. Both Lagrange and de Foncenex published several papers in early volumes of the *Miscellanea*, with de Foncenex crediting Lagrange for much of the inspiration behind his own work. Delambre argued that Lagrange provided de Foncenex with far more than inspiration, and it is true that de Foncenex's analytic style is strongly reminiscent of Lagrange. It is also true that de Foncenex did not live up to the mathematical promise demonstrated in his early work, although he was perhaps sidetracked from a mathematical career after being named head of the navy by the King of Sardinia as a result of his early successes in the *Miscellanea*.

In his earliest paper, *Reflexions sur les Quantités Imaginaire* [**7**], de Foncenex focused his attention on "the nature of imaginary roots" within the debate concerning logarithms of negative quantities. In particular, de Foncenex wished to reconcile Euler's "incontestable calculations" proving that negative numbers have imaginary logarithms with an argument from Bernoulli that opposed this conclusion on grounds involving the continuity of the hyperbola (whose quadrature defines logarithms) at infinity. The analysis that de Foncenex developed of this problem led him to consider the relation between the circle and the equilateral hyperbola—exactly the same analogy pursued by Lambert.

In his 1761 *Mémoire*, Lambert fully credited de Foncenex with having shown how the affinity between the circular trigonometric functions and the hyperbolic trigonometric functions can be "seen in a very simple and direct fashion by comparing the circle and the equilateral hyperbola with the same center and same diameter." De Foncenex himself went no further in exploring "this affinity" than to conclude that, since $\sqrt{x^2 - r^2} = \sqrt{-1}\sqrt{r^2 - x^2}$, "the circular sectors and hyperbolic [sectors] that correspond to the same abscissa are always in the ratio of 1 to $\sqrt{-1}$." It is this use of an imaginary ratio to pass from the circle to the hyperbola Lambert seemed intent on avoiding.

Lambert returned to this theme one final time in Section 88 of the *Mémoire*. In another classic example of 18th-century analysis, Lambert first remarked that "one can easily find by using the differential formulas of Section 75," that

$$v = \operatorname{tang}\phi - \frac{1}{3}\operatorname{tang}\phi^3 + \frac{1}{5}\operatorname{tang}\phi^5 - \frac{1}{7}\operatorname{tang}\phi^7 + \text{etc.}$$

$$\operatorname{tang}\phi = u - \frac{1}{3}u^3 + \frac{2}{15}u^5 - \frac{17}{315}u^7 + \text{etc.}$$

"By substituting the value of the second of these series into the first ... and reciprocally" (but again with details omitted), Lambert obtained the following two series:

$$v = u - \frac{2}{3}u^3 + \frac{2}{3}u^5 - \frac{244}{315}u^7 + \text{etc.} \tag{2}$$

$$u = v + \frac{2}{3}v^3 + \frac{2}{3}v^5 + \frac{244}{315}v^7 + \text{etc.}$$

where (switching from previous usage) u equals twice the area of the circular sector and v equals twice the area of the hyperbolic sector. Finally, Lambert obtained the sought-after relation by noting that substitution of $u = v\sqrt{-1}$ into series (2) will yield $v = u\sqrt{-1}$.

In (semi)-modern notation, we can represent Lambert's results as $\tanh(v\sqrt{-1}) = \tan(u\sqrt{-1})$ and $\tanh(u) = \tan(v)$. Having thus established that imaginary hyperbolic sectors correspond to imaginary circular sectors, and similarly for real sectors, Lambert closed his 1761 *Mémoire*. The next scene examines how he later pursued a new plot line suggested by this analogy: the use of hyperbolic functions to replace circular functions in the solution of certain problems.

Act II, Scene II: The reappearance of hyperbolic functions in Lambert

Lambert returned to the development of his "transcendental logarithmic functions" and their similarities to circular trigonometric functions in his 1768 paper *Observations trigonometriques* [**15**]. In this treatment, a typical point on the hyperbola is called q. Letting ϕ denote the angle qCQ in FIGURE 5, Lambert first remarked that $\text{tang}\,\phi = MN/MC = qp/pC$. Because $MN/MC = \sin\phi/\cos\phi$ and $qp/pC = \sin\text{hyp}\,\phi/\cos\text{hyp}\,\phi$, one has the option of using either the circular tangent function or the hyperbolic tangent functions for the purpose of analyzing triangle qCP. *Note that the notation and terminology used here are Lambert's own!* Lambert himself commented that, in view of the analogous parameterizations that are possible for the circle and the hyperbola, there is "nothing repugnant to the original meaning" of the terms "sine" and "cosine" in the use of the terms "hyperbolic sine" and "hyperbolic cosine" to denote the abscissa and ordinate of the hyperbola. Although Lambert's notation for these functions differed from our current convention, the hyperbolic functions had now become fully-fledged players in their own right, complete with names and notation suggestive of their relation to the circular trigonometric functions.

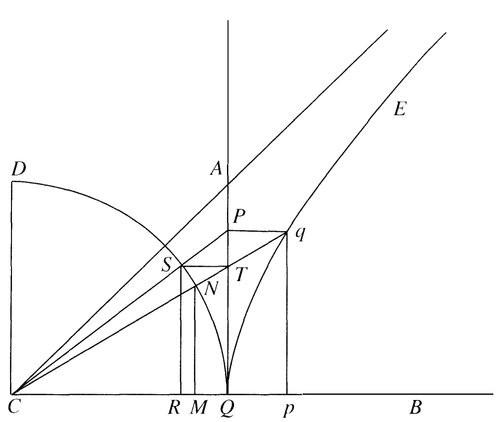


Figure 5 Diagram from Lambert's *Observations trigonometriques* [**15**]

The development of the hyperbolic functions in this paper included an extensive list of sum, difference, and multi-angle identities that are, as Lambert remarked, easily derived from the formulas $\sin\text{hyp}\, v = (e^v + e^{-v})/2$, $\cos\text{hyp}\, v = (e^v - e^{-v})/2$. Of greater importance to Lambert's immediate purpose was the table of values he constructed for certain functions of "the transcendental angle ω." In particular, the transcendental angle ω, defined as angle PCQ in FIGURE 5 and related to the common angle ϕ via the relation $\sin\omega = \text{tang}\,\phi = \text{tang hyp}\,\phi$, served Lambert as a means to pass from circular functions to hyperbolic functions. (The transcendental angle associated with ϕ is also known as the *hyperbolic amplitude of* ϕ after Hoüel and the *longitude* after Guderman.) For values of ω ranging from 1° to 90° in increments of 1 degree, Lambert's table included values of the hyperbolic sector, the hyperbolic sine and its logarithm, the hyperbolic cosine and its logarithm, as well as the tangent of the corresponding common angle and its logarithm. By replacing circular functions by hyperbolic functions, Lambert used these functions to simplify the computations required to determine the angle measures and the side lengths of certain triangles.

The triangles that Lambert was interested in analyzing with the aid of the hyperbolic functions arise from problems in astronomy in which one of the celestial bodies is below the horizon. It has since been noted that such problems can be solved using formulae from spherical trigonometry with arcs that are pure imaginaries. This is an intriguing observation since elsewhere (in his work on noneuclidean geometry), Lambert speculated on the idea that a sphere of imaginary radius might reflect the geometry of "the acute angle hypothesis." The acute angle hypothesis is one of three possibilities for the two (remaining) similar angles α, β of a quadrilateral assumed to have two right angles and two congruent sides: (1) angles α, β are right; (2) angles α, β are obtuse; and (3) angles α, β are acute. Girolamo Saccheri (1667–1733) introduced this quadrilateral in his *Euclides ab omi naevo vindiactus* of 1733 as an element of his efforts to prove Euclid's Fifth Postulate by contradiction. Both Saccheri and Lambert believed they could dispense with the obtuse angle hypothesis. Lambert's speculation about the acute angle hypothesis was the result of his inability to reject the acute angle hypothesis.

It is worth emphasizing, however, that Lambert himself never put an imaginary radius into the formulae of spherical trigonometry in any of his published works. The triangles he treated are real triangles with real-valued arcs and real-valued sides. As noted by historian Jeremy Gray [**9**, pp. 156–158], the ability to articulate clearly the notion of "geometry on a sphere of imaginary radius" was not yet within the grasp of mathematicians in the age of Euler. Gray argues convincingly that the development of analysis by Euler, Lambert, and other 18th-century mathematicians was, nevertheless, critical for the 19th-century breakthroughs in the study of noneuclidean geometry. By providing a language flexible enough to discuss geometry in terms other than those set forth by Euclid, analytic formulae allowed for a reformulation of the problem and the recognition that a new geometry for space was possible. Although rarely mentioned in today's calculus texts, the explicit connection eventually made by Beltrami in his 1868 paper, linking the hyperbolic functions to the noneuclidean geometry of an imaginary sphere, is yet another intriguing use for hyperbolic functions that is surely as tantalizing as the oft-cited catenary curve.

Flashback: Hyperbolic functions in Riccati Although Lambert's primary reason for considering hyperbolic functions in 1768 was to simplify calculations involved in solving triangles, Lambert clearly realized that there was no need to define new functions for this purpose; tables of logarithms of appropriate trigonometric values could instead be used to serve the same end. But, he argued, this was only one possible use for the hyperbolic trigonometric functions in mathematics. The only example he cited in this regard was the simplification of solution methods for equations. Lambert did not elaborate on this idea beyond noting that the equation $0 = x^2 - 2a\cos\omega \cdot x + a^2$ is equivalent to the equation $0 = x^2 - 2a\text{coshyp}\psi \cdot x + a^2$ for an appropriately defined angle ψ. He did, however, cite an investigation of this idea that had already appeared in the work of another 18th-century mathematician: Vincenzo de Riccati.

Vincenzo de Riccati was born on January 11, 1707, the second son of Jacopo Riccati for whom the Riccati equation in differential equations is named. Riccati (the son) received his early education at home and from the Jesuits. He entered the Jesuit order in 1726 and taught or studied in various locations, including Piacenza, Padua, Parma, and Rome. In 1739, Riccati moved to Bologna, where he taught mathematics in the College of San Francesco Saverio until Pope Clement XIV suppressed the Society of Jesus in 1773. Riccati then returned to his family home in Treviso, where he died on January 17, 1775.

Riccati first treated hyperbolic functions in his two-volume *Opuscula ad res physicas et mathematicas pertinentium* (1757–1762) **[19]**. In this work, Riccati employed a hyperbola to define functions that he referred to as "sinus hyperbolico" and "cosinus hyperbolico," doing so in a manner analogous to the use of a circle to define the functions "sinus circulare" and "cosinus circulare." Taking u to be the quantity given by twice the area of the sector ACF divided by the length of the segment CA (whether in the circle or the hyperbola of FIGURE 6), Ricatti defined the sine and cosine of the quantity u to be the segments GF and CG of the appropriate diagram. Although Ricatti did not explicitly assume either a unit circle or an equilaterial hyperbola, his definitions are equivalent to that of Lambert (and our own) in that case. In *Opusculum* IV of Volume I, Ricatti derived several identities of his hyperbolic sine and cosine, applying these to the problem of determining roots of certain equations, especially cubics. Riccati also determined the series representations for the *sinus* and *cosinus hyperbolicos*. These latter results, which appeared in *Opusculum* VI in volume I, were earlier communicated by Riccati to Josepho Suzzio in a letter dated 1752.

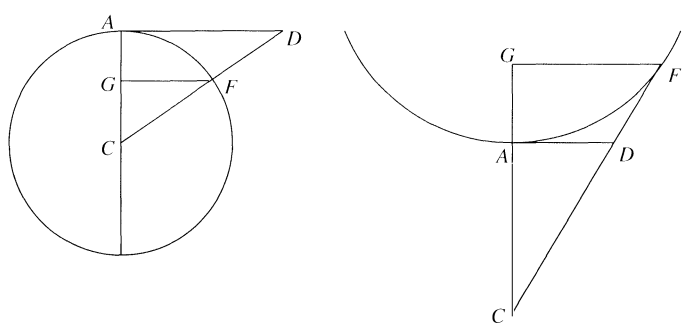


Figure 6 Diagrams rendered from Riccati's *Opuscula*

In Riccati's *Institutiones analyticae* (1765–1767) **[20]**, written collaboratively with Girolamo Saldini, he further developed the theory of the hyperbolic functions, including the standard addition formulas and other identities for hyperbolic functions, their derivatives and their relation to the exponential function (already implicit in his *Opuscula*).

Reprise: Giving credit where credit is due While some of the ideas in Riccati's *Institutiones* of 1765–1767 also appeared in Lambert's 1761 *Mémoire*, this author knows of no evidence to suggest that Riccati was building on Lambert's work. The publication dates of his earlier work suggest that Riccati was familiar with the analogy between the circular and the hyperbolic functions some time earlier than Lambert came across the idea, and certainly no later. Conversely, even though Riccati's earliest work was published several years before Lambert's 1761 *Mémoire*, it appears that Lambert was unfamiliar with Riccati's work at that time. Certainly, the motivations of the two for introducing the hyperbolic functions appear to have been quite different. Furthermore, Lambert appears to have been scrupulous in giving credit to colleagues when drawing on their work, as in the case of de Foncenex. In fact, Lambert credited Ricatti with developing the terminology "hyperbolic sine" and "hyperbolic cosine" when he used these names for the first time in his 1768 *Observations trigonometriques*. It thus appears that it was only these new names—and perhaps the idea of using these functions to solve equations—that Lambert took from Riccati's work, finding them to be suitable nomenclature for mathematical characters whom he had already developed within a story line of his own creation.

Despite the apparent independence of their work, the fact remains that Riccati did have priority in publication. Why then is Lambert's name almost universally mentioned in this context, with Riccati receiving little or no mention? Histories of mathematics written in the 19th and early 20th centuries suggest this tendency to overlook Riccati's work is a relatively recent phenomenon. Von Braunmühl [**22**, pp. 133–134], for example, has the following to say in his 1903 history of trigonometry:

> In fact, Gregory St. Vincent, David Gregory and Craig through the quadrature of the equilateral hyperbola, erected the foundations [for the hyperbolic functions], even if unaware of the fact, Newton touched on the parallels between the circle and the equilateral hyperbola, and de Moivre seemed to have some understanding that, by substituting the real for the imaginary, the role of the circle is replaced by the equilateral hyperbola. Using geometric considerations, Vincenzo Riccati (1707–1775) was the first to found the theory of hyperbolic functions, as was recognized by Lambert himself. *(Author's translation.)*

Although the amount of recognition that Lambert afforded Riccati may be overestimated here, it is interesting that von Braunmühl then proceeded to discuss Lambert's work on hyperbolic functions in detail, with no further mention of Riccati, remarking that:

> This [hyperbolic function] theory is only of interest to us in so far as it came into use in the treatment of trigonometric problems, as was first opened up by Lambert. *(Author's translation.)*

It would thus appear that the motivation Lambert assigned to the hyperbolic functions was more central to mathematical interests as they evolved thereafter, even though his interests often fell outside the mainstream of his own century. The fact that Lambert's mathematical works, especially those on noneuclidean geometry, were studied by his immediate mathematical successors offers support for this idea, as does the wider availability of Lambert's works today. Besides being more widely available, Lambert's work is written in notation—and languages!—that are more familiar to today's scholars than that of Riccati. This alone makes it easier to tell Lambert's story in more detail, just as we have done here.

Epilogue

And what of the physical applications for which the hyperbolic functions are so useful? Although neither Lambert nor Riccati appear to have studied these connections, they were known by the late 19th century, as evidenced by the publication of hyperbolic function tables and manuals for engineers in that period. Yet even as late as 1849, we hear Augustus De Morgan [**3**, p. 66], declare:

> The system of trigonometry, from the moment that $\sqrt{-1}$ is introduced, always presents an incomplete and one-sided appearance, unless the student have in his mind for comparison *(though it is rarely or never wanted for what is called use)*, another system [hyperbolic trigonometry] in which the there-called sines and cosines are real algebraic quantities. (Emphasis added.)

While De Morgan's perspective offers yet another intriguing reason to study hyperbolic trigonometry, usefulness in solving problems (mathematical or physical) did not appear to concern him. This delay between the development of the mathematical machinery and its application to physical problems serves as a gentle reminder that the physical applications we sometimes cite as the raison d'être for a mathematical idea may only become visible with hindsight. Yet even Riccati's and Lambert's own uses for hyperbolic trigonometry went unacknowledged by De Morgan—an even stronger reminder of how quickly mathematics changed in the 19th century, and how greatly today's mathematics classroom might be enriched by remembering the mathematics of the age of Euler.

Acknowledgments. Many thanks to my Calculus II students from Spring 2000 for asking questions about the history of the hyperbolic functions, and to the organizers of the 2001 New Orleans MAA Contributed Session "Mathematics in the Age of Euler" for encouraging me to publish the answers. The editorial assistance and suggestions of William Dunham, George W. Heine, and the anonymous referees of this MAGAZINE are also greatly appreciated.

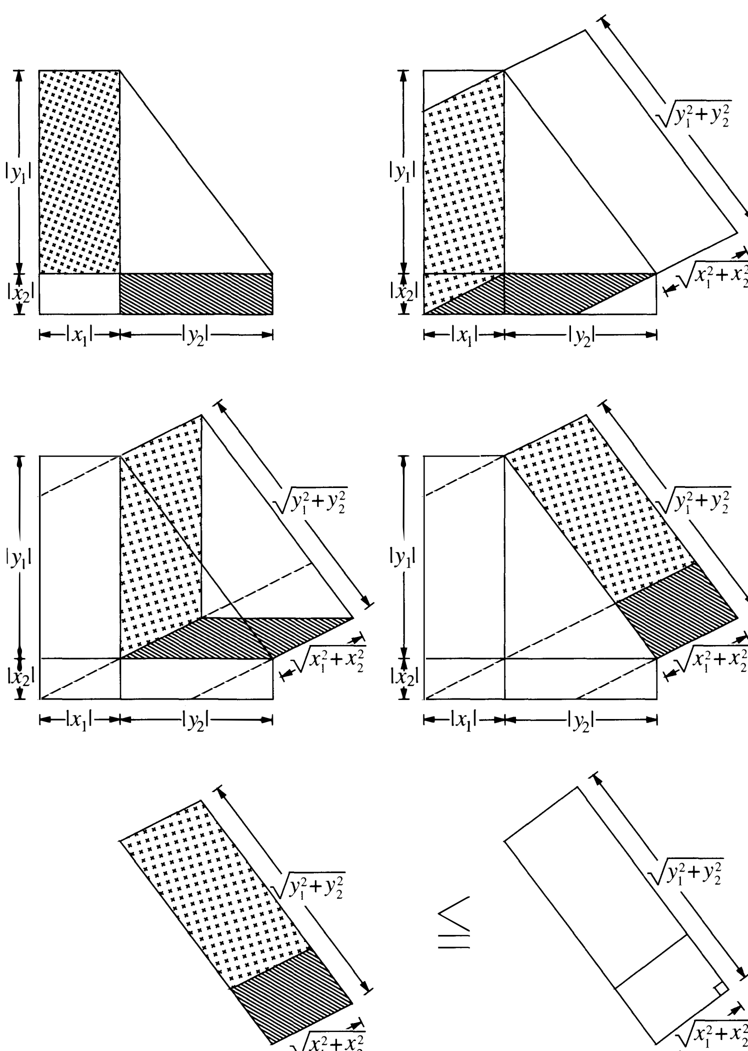
REFERENCES

1. Johannis Bernoulli, Solutio problematis funicularii, *Acta Eruditorum* (June 1691), 274–276.
2. J. B. J. Delambre, Notice sur a vie et les ouvrages de M. le Comte J. L. Lagrange, 1812. Reprinted in *Oeuvre de Lagrange* **I**, Gauthier - Villars, Paris, 1816, ix–li.
3. Augustus De Morgan, *Trigonometry and Double Algebra*, Metcalf and Palmer, Cambridge, 1849.
4. René Descartes, *The Geometry*, translated by David Eugene Smith and Marcia L. Latham, Dover, New York, 1954.
5. Leonhardi Euleri, *Methodus Inveniendi Lineas Curvas Maximi Minimive Proprietate Gaudentes*, 1744. Reprinted in *Leonhardi Euleri Opera Omnia, Series Prima* **24**, B. G. Teubneri, Leipzig and Berlin, 1952.
6. ———, *Introductio in Analysin Infinitorum, Tomus Primus*, 1748. Reprinted in *Leonhardi Euleri Opera Omnia, Series Prima* **8**, B. G. Teubneri, Leipzig and Berlin, 1922.
7. Daviet de Foncenex, Réflexions sur les quantités imaginaries, *Miscellanea philosophica-mathematica Societatis Privatae Taurinensis*, **I** (1759), 113–146.
8. Galileo, *Two New Sciences*, tr. Stillman Drake, U. of Wisconsin Press, Madison, 1974, p. 143.
9. J. J. Gray, *Ideas of Space: Euclidean, Non-Euclidean and Relativistic*, Clarendon Press, Oxford, 1989.
10. J. J. Gray and L. Tilling, Johann Heinrich Lambert, Mathematician and Scientist, 1728–1777, *Hist. Math.* **5** (1978), 13–41.
11. Christiani Huygens, Dynaste Zulichemii, solutio problematis funicularii, *Acta Eruditorum* (June 1691), 281–282.
12. Victor Katz, *A History of Mathematics: An Introduction*, Harper Collins, New York, 1993.
13. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Pr., New York, 1978.
14. Johann Lambert, Mémoire sur quelques propriétés remarquable des quantités transcendantes circulaires et logarithmiqes, *Mémoires de l'Académies des Sciences de Berlin* **17**, (1761, 1768), 265–322. Reprinted in *Iohannis Henrici Lamberti Opera mathematica*, **2**, Orell Füssli, Turici, 1948, 112–159.
15. ———, Observations Trigonométriques, *Mémoires de l'Académies des Sciences de Berlin* **24** (1768), 1770, 327–354. Reprinted in *Iohannis Henrici Lamberti Opera mathematica*, **2**, Orell Füssli, Turici, 1948, 245–269.
16. Gottfried Leibniz, Solutio problematic catenarii, *Acta Eruditorum* (June 1691), 277–281.
17. Mark McKinzie and Curtis Tuckey, Higher Trigonometry, Hyperreal Numbers, and Euler's Analysis of Infinites, this MAGAZINE **74**:5 (2001), 339–367.
18. A. Natucci, Vincenzo Riccati, *Biographical Dictionary of Mathematicians*, Macmillan, New York, 1991, 2126–2127.

19. Vincenzo Riccati, *Opuscula ad res physicas et mathematicas pertinentium*, **I**, Apud Laelium a Vulpe Instituti Scientiarum Typographum, Bononiae, 1757–1762.
20. Vincenzo Riccati (with G. Saladini), *Institutiones analyticae*, Apud Laelium a Vulpe Instituti Scientiarum Typographum, Bononiae, 1765–1767.
21. C. Scriba, Johann Heinrich Lambert, *Biographical Dictionary of Mathematicians*, Macmillan, New York, 1991, 1323–1328.
22. A. Von Braunmühl, *Vorlesungen über geschicten der trigonometrie*, Druckund Verlag von B. G. Teubneri, Leipzig, 1903.

Proof Without Words: Cauchy-Schwarz Inequality

$$|x_1y_1 + x_2y_2 \leq| \, |x_1| \cdot |y_1| + |x_2| \cdot |y_2| \leq \sqrt{x_1^2 + x_2^2} \cdot \sqrt{y_1^2 + y_2^2}$$



—CLAUDI ALSINA
UNIVERSITAT POLITÈCNICA DE CATALUNYA
08028 BARCELONA, SPAIN

Simpson Symmetrized and Surpassed

DANIEL J. VELLEMAN
Amherst College
Amherst, MA 01002
djvelleman@amherst.edu

Simpson's rule is a well-known numerical method for approximating definite integrals. It is named after Thomas Simpson, who published it in 1743, although it was known already more than a century before that. Bonaventura Cavalieri gave a geometric version of Simpson's rule in 1639, and James Gregory published the rule in 1668. Others who published not only Simpson's rule but also more general formulas before Simpson's publication in 1743 include Isaac Newton, Roger Cotes, and James Stirling [**4**, p. 77].

Many calculus textbooks state Simpson's rule in its composite form, which says that

$$\int_a^b f(x)\,dx \approx \Delta x\left[\frac{1}{3}f(x_0)+\frac{4}{3}f(x_1)+\frac{2}{3}f(x_2)+\frac{4}{3}f(x_3)+\cdots\right.$$
$$\left.+\frac{2}{3}f(x_{n-2})+\frac{4}{3}f(x_{n-1})+\frac{1}{3}f(x_n)\right],$$

where n is a positive even integer, $\Delta x = (b-a)/n$, and $x_i = a + i\,\Delta x$ for $0 \le i \le n$.

Simpson's rule is surprisingly accurate. For example, although it is based on approximating the function f on various intervals with quadratic polynomials, it is exactly correct even if f is a cubic polynomial. However, the asymmetric treatment of the even- and odd-numbered sample points that results from the alternation of the coefficients 4/3 and 2/3 seems counterintuitive. The sample points are evenly spaced between a and b, so once one gets away from the endpoints of the interval, every sample point looks very much like every other one. Why should adjacent sample points be treated so differently? Others have raised the same issue before. For example, Roger Pinkham [**8**, p. 92] argues that "... the function evaluations in the middle of the interval are on an equal footing. One feels that they should be treated evenhandedly."

In this paper I will show that symmetrizing the treatment of even- and odd-numbered sample points in Simpson's rule can lead to more accurate approximate integration formulas. These formulas will still be based on approximating f on intervals with quadratic polynomials, and they will still be exact for cubic polynomials. However, the error bounds will be smaller than the error bound for Simpson's rule, and all coefficients except for a few at the beginning and end will be equal to 1.

It is not my intention in this paper to study numerical integration in general. Rather, I will focus on the limited topic of Simpson-like numerical integration rules that are based on quadratic approximation. My question is not whether Simpson's rule is the best way to approximate definite integrals, but rather whether Simpson's rule, with its asymmetric treatment of even- and odd-numbered sample points, is the best way to employ quadratic approximations in numerical integration.

A first attempt

It will be helpful to begin by reviewing briefly the derivation of Simpson's rule for approximating $\int_a^b f(x)\,dx$. The first step of this derivation is to divide the interval

$[a, b]$ into n equal subintervals of width $\Delta x = (b-a)/n$, for some positive even integer n. The dividing points are $x_i = a + i\Delta x$, for $0 \le i \le n$. For each even integer i, $0 \le i \le n-2$, we then approximate f on the interval $[x_i, x_{i+2}]$ with a quadratic polynomial q such that $q(x_i) = f(x_i)$, $q(x_{i+1}) = f(x_{i+1})$, and $q(x_{i+2}) = f(x_{i+2})$. It is not hard to show that there is exactly one such polynomial q, and it is given by the formula

$$q(x) = A(x - x_{i+1})^2 + B(x - x_{i+1}) + C, \tag{1}$$

where

$$A = \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{2\Delta x^2}, \qquad B = \frac{f(x_{i+2}) - f(x_i)}{2\Delta x}, \qquad C = f(x_{i+1}).$$

Finally, we approximate the integral of f over the interval $[x_i, x_{i+2}]$ with the integral of q over the same interval, which we evaluate using the substitution $u = x - x_{i+1}$:

$$\begin{aligned}\int_{x_i}^{x_{i+2}} f(x)\,dx \approx \int_{x_i}^{x_{i+2}} q(x)\,dx &= \int_{-\Delta x}^{\Delta x} Au^2 + Bu + C\,du \\ &= \Delta x\left[\frac{1}{3}f(x_i) + \frac{4}{3}f(x_{i+1}) + \frac{1}{3}f(x_{i+2})\right]. \end{aligned} \tag{2}$$

(See FIGURE 1.) Summing these approximations yields Simpson's rule.

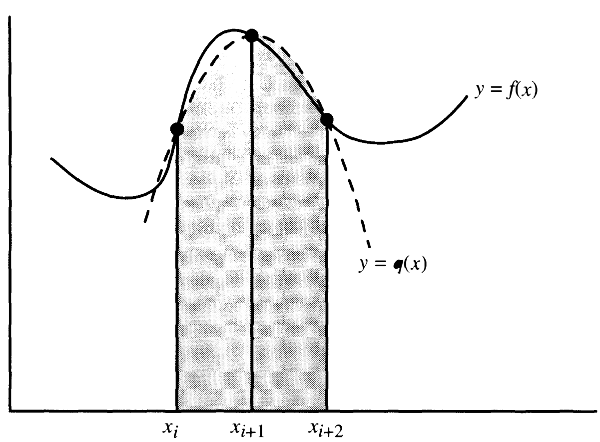


Figure 1 Approximating $\int_{x_i}^{x_{i+2}} f(x)\,dx$

We define the *error* in any approximation of the integral of f on an interval to be the exact value of the integral minus the approximation; thus, the error is positive if the approximation is too small, and negative if it is too large. It is clear from the derivation that Simpson's rule is exactly correct if f is a quadratic polynomial. Surprisingly, it is also exactly correct for cubics. To understand why, it is helpful to break the basic Simpson's rule approximation (2) into left and right halves. With q chosen as before, it is easy to compute that

$$\int_{x_i}^{x_{i+1}} q(x)\,dx = \Delta x\left[\frac{5}{12}f(x_i) + \frac{2}{3}f(x_{i+1}) - \frac{1}{12}f(x_{i+2})\right] \tag{3}$$

and

$$\int_{x_{i+1}}^{x_{i+2}} q(x)\,dx = \Delta x\left[-\frac{1}{12}f(x_i) + \frac{2}{3}f(x_{i+1}) + \frac{5}{12}f(x_{i+2})\right]. \tag{4}$$

Notice that these formulas sum to the formula in (2).

Now consider any cubic polynomial $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$. Let $h(x) = a_3(x - x_i)(x - x_{i+1})(x - x_{i+2})$. Then it is not hard to see that $f - h$ is a quadratic polynomial, and it agrees with f at x_i, x_{i+1}, and x_{i+2}, so it must be the function q in equations (2)–(4). It follows that the error in using the integral of q to approximate the integral of f on any interval can be found by integrating $f - q = h$ on that interval. It is now clear from the symmetry of FIGURE 2 that the errors in using (3) and (4) to approximate the integrals of f on the intervals $[x_i, x_{i+1}]$ and $[x_{i+1}, x_{i+2}]$ are equal in magnitude but have opposite sign. Indeed, it is straightforward to calculate that these errors are $\pm a_3\Delta x^4/4$. These errors therefore cancel each other out, making the basic Simpson's rule approximation (2) exactly correct for f. (Kenneth Supowit uses a similar approach to prove a generalization of this fact to all Newton-Cotes formulas of even degree [**9**].)

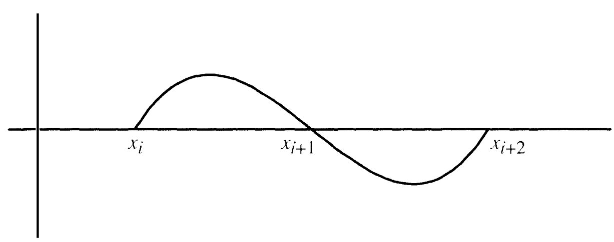


Figure 2 The graph of $y = h(x) = f(x) - q(x)$ when f is a cubic polynomial

In fact, even if f is not a cubic polynomial, it is often the case that the errors in the left and right halves of the approximation (2) cancel to some extent, although not exactly. For example, such cancellation can be seen in FIGURE 1. Intuitively, this cancellation helps to explain the high degree of accuracy of Simpson's rule.

The derivation of Simpson's rule shows that the source of the asymmetry in the coefficients of the even- and odd-numbered sample points is the fact that the basic Simpson's rule approximation (2) is used to approximate $\int_{x_i}^{x_{i+2}} f(x)\,dx$ only for *even* i. This observation suggests a simple approach to symmetrizing Simpson's rule that was proposed by G. O. Peters and C. E. Maley [**6**]. Their idea is to apply (2) for *every* i between 0 and $n - 2$, rather than just for even i. The intervals used in these approximations overlap and cover the interval $[a, b]$ twice, except for the intervals $[x_0, x_1]$ and $[x_{n-1}, x_n]$, which are covered only once. Thus, summing these approximations, and then adding an additional approximation of the integral of f on each of the intervals $[x_0, x_1]$ and $[x_{n-1}, x_n]$, yields an approximation of twice the desired integral. Peters and Maley use (3), the left half of the basic Simpson's rule approximation, for the additional approximation of the integral on $[x_0, x_1]$, and the right half (4) for $[x_{n-1}, x_n]$. Finally, dividing the sum of all of these approximations by two yields the following symmetrized version of Simpson's rule:

$$\int_a^b f(x)\,dx \approx \Delta x\left[\frac{3}{8}f(x_0) + \frac{7}{6}f(x_1) + \frac{23}{24}f(x_2) + f(x_3) + f(x_4) + \cdots \right.$$
$$\left. + f(x_{n-3}) + \frac{23}{24}f(x_{n-2}) + \frac{7}{6}f(x_{n-1}) + \frac{3}{8}f(x_n)\right].$$

As G. M. Phillips has observed [**7**], Peters and Maley's formula is the same as Gregory's rule of order two, also sometimes called the Lacroix rule. Gregory's rules are usually derived by a different method, involving the use of the Euler-Maclaurin summation formula to determine correction terms that are added to the trapezoid rule. J. M. De Villiers has also given a derivation of the Lacroix rule using quadratic approximation [**3**].

As promised, all coefficients in this symmetrized version of Simpson's rule except for a few at the beginning and end are equal to 1. The symmetrized rule is also exactly correct for cubic polynomials, since it consists of equal numbers of left and right halves of the basic Simpson's rule approximation, and the errors in these approximations cancel out for cubic polynomials. And it has the modest advantage that it can also be used if n is odd.

Unfortunately, this symmetrized Simpson's rule also has a serious disadvantage: Its error bound is larger than the error bound for Simpson's rule! If f is continuous on $[a, b]$, and $f^{(4)}(x)$ is defined and $|f^{(4)}(x)| \leq M$ for all $x \in (a, b)$, then the magnitude of the error in Simpson's rule is at most

$$\frac{n\Delta x^5}{180}M = \frac{(b-a)^5}{180n^4}M. \tag{5}$$

(We will see the derivation of this shortly.) But it turns out that the magnitude of the error in Peters and Maley's symmetrized rule can be as large as

$$\left(\frac{19n}{720} - \frac{1}{24}\right)\Delta x^5 M = \left(\frac{19}{4} - \frac{15}{2n}\right) \cdot \frac{(b-a)^5}{180n^4}M.$$

In particular, this is the magnitude of the error in the case of the function $f(x) = x^4$. For large n, this error is approximately $19/4$ times the error in Simpson's rule. (Peters and Maley do not provide an estimate of the error in their approximation. The *Mathematical Reviews* entry for their paper (MR 38 #4032) does provide an error estimate, but it is incorrect.)

Roger Pinkham takes a similar approach to symmetrizing Simpson's rule [**8**], although he deals with the intervals $[x_0, x_1]$ and $[x_{n-1}, x_n]$ differently. Pinkham's approximation is more accurate than Peters and Maley's, but this improvement comes at the price of using an additional sample point in each of the intervals $[x_0, x_1]$ and $[x_{n-1}, x_n]$. When compared to Simpson's rule with the same number of sample points, Pinkham's rule is slightly less accurate, although the relative difference is small for large n. Thus, the rules given by Peters, Maley, and Pinkham are Simpson symmetrized, but not yet Simpson surpassed.

A better way

Why is Simpson's rule so accurate? It will turn out that an examination of the proof of the error bound (5) for Simpson's rule will lead us to a symmetrized version of Simpson's rule with a lower error bound. We will follow Apostol's proof [**1**, pp. 605–609].

Assume that f is continuous on $[a, b]$, and that $f^{(4)}(x)$ is defined and $|f^{(4)}(x)| \leq M$ for all $x \in (a, b)$. To prove the error bound (5), we begin by considering the basic Simpson's rule approximation (2) for $\int_{x_i}^{x_{i+2}} f(x)\,dx$. It is not hard to show that there is a cubic polynomial g such that $g(x_i) = f(x_i)$, $g(x_{i+1}) = f(x_{i+1})$, $g(x_{i+2}) = f(x_{i+2})$, and $g'(x_{i+1}) = f'(x_{i+1})$. Since g is a cubic polynomial, Simpson's rule is exact for g, and therefore

$$\int_{x_i}^{x_{i+2}} g(x)\,dx = \Delta x\left[\frac{1}{3}g(x_i) + \frac{4}{3}g(x_{i+1}) + \frac{1}{3}g(x_{i+2})\right]$$
$$= \Delta x\left[\frac{1}{3}f(x_i) + \frac{4}{3}f(x_{i+1}) + \frac{1}{3}f(x_{i+2})\right].$$

It follows that the magnitude of the error in the approximation (2) is

$$\left|\int_{x_i}^{x_{i+2}} f(x)\,dx - \int_{x_i}^{x_{i+2}} g(x)\,dx\right| \le \int_{x_i}^{x_{i+2}} \left|f(x) - g(x)\right| dx. \tag{6}$$

Thus, to estimate the error in (2), we must investigate the size of $|f(x) - g(x)|$.

Notice that the function $f - g$ takes on the value 0 at x_i, x_{i+1}, and x_{i+2}, and furthermore its derivative is also 0 at x_{i+1}. Perhaps the simplest function with these properties is the function w defined as follows:

$$w(x) = (x - x_i)(x - x_{i+1})^2(x - x_{i+2}).$$

One might hope that there is some relationship between the functions $f - g$ and w, and it turns out that there is:

LEMMA. For every $x \in [x_i, x_{i+2}]$ there is some $c_x \in (x_i, x_{i+2})$ such that

$$f(x) - g(x) = \frac{f^{(4)}(c_x)}{4!}w(x). \tag{7}$$

Sketch of proof: The lemma clearly holds for $x = x_i$, $x = x_{i+1}$, and $x = x_{i+2}$, since in these cases both sides of (7) are 0. For other values of x, the proof involves applying Rolle's theorem repeatedly to the function $h(t) = w(x)(f(t) - g(t)) - w(t)(f(x) - g(x))$. For details, see Apostol [**1**, p. 608, equation (15.42)]. ∎

Since $|f^{(4)}(c)| \le M$ for all $c \in (x_i, x_{i+2})$, the lemma implies that for all $x \in [x_i, x_{i+2}]$,

$$\left|f(x) - g(x)\right| \le \frac{M}{4!}\left|w(x)\right|.$$

Thus, by (6), the error in (2) is at most

$$\int_{x_i}^{x_{i+2}} \left|f(x) - g(x)\right| dx \le \frac{M}{4!}\int_{x_i}^{x_{i+2}} \left|w(x)\right| dx. \tag{8}$$

To complete the calculation of the error bound, we observe that $w(x) \le 0$ on $[x_i, x_{i+2}]$, so $|w(x)| = -w(x)$ on this interval, and then we integrate using the substitution $u = x - x_{i+1}$:

$$\int_{x_i}^{x_{i+2}} \left|w(x)\right| dx = \int_{x_i}^{x_{i+2}} -w(x)\,dx = \int_{-\Delta x}^{\Delta x} u^2\Delta x^2 - u^4\,du = \frac{4\Delta x^5}{15}.$$

Plugging this into (8), we find that the magnitude of the error in the basic Simpson's rule approximation (2) is at most

$$\frac{M}{4!}\cdot\frac{4\Delta x^5}{15} = \frac{\Delta x^5}{90}M. \tag{9}$$

Since Simpson's rule is a sum of $n/2$ of these basic approximations, the total error in Simpson's rule is at most $n/2$ times the error bound (9), which gives us the error bound (5).

This proof shows that the size of the error bound for Simpson's rule is determined by the integral of the function $|w(x)|$. FIGURE 3 shows the graph of w, and what is most striking in this graph is that, because of the double root at x_{i+1}, the value of $|w(x)|$ is fairly small for x near the middle of the interval $[x_i, x_{i+2}]$, with the largest values occurring close to the endpoints of the interval. It appears that, in some sense, most of the error in (2) comes from the beginning and end of the interval, with less error coming from near the middle. We might say that the middle of the interval is the "sweet spot" of the approximation (2). This observation will be the motivation for our improvements on Simpson's rule. In our new approximations of $\int_a^b f(x)\,dx$, we will continue to fit quadratic polynomials to f on intervals of width $2\Delta x$, but we will only integrate these quadratic polynomials over the middle half of each interval.

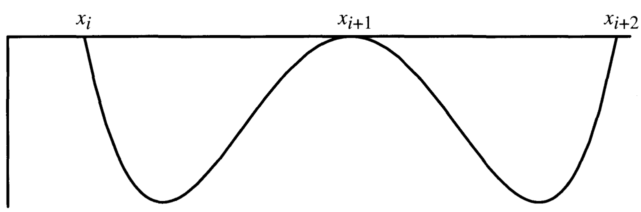


Figure 3 The graph of $y = w(x)$

As before, in order to approximate $\int_a^b f(x)\,dx$, we begin by dividing the interval $[a, b]$ into n equal subintervals of width $\Delta x = (b-a)/n$, at points $x_i = a + i\Delta x$, $0 \le i \le n$. (There will be no need to assume that n is even for this approximation.) It will be convenient to introduce the additional notation $x_{1/2} = (x_0 + x_1)/2$, $x_{3/2} = (x_1 + x_2)/2$, and so on. For each integer i, $0 \le i \le n-2$, we again find the unique quadratic polynomial q that agrees with f at x_i, x_{i+1}, and x_{i+2}; the formula for $q(x)$ is given by (1). However, we only integrate f and q over the interval $[x_{i+1/2}, x_{i+3/2}]$, again using the substitution $u = x - x_{i+1}$ to evaluate the integral of q:

$$\int_{x_{i+1/2}}^{x_{i+3/2}} f(x)\,dx \approx \int_{x_{i+1/2}}^{x_{i+3/2}} q(x)\,dx = \int_{-\Delta x/2}^{\Delta x/2} Au^2 + Bu + C\,du$$
$$= \Delta x\left[\frac{1}{24}f(x_i) + \frac{11}{12}f(x_{i+1}) + \frac{1}{24}f(x_{i+2})\right]. \qquad (10)$$

(See FIGURE 4.)

Before continuing, let us stop to see how accurate the approximation in (10) is. We will assume, as before, that f is continuous on $[a, b]$, and that $f^{(4)}(x)$ is defined and $|f^{(4)}(x)| \le M$ for all $x \in (a, b)$. It is easy to verify, by an argument similar to the one given earlier for Simpson's rule, that (10) is exact if f is a cubic polynomial. As a result, we can imitate our derivation of the error in Simpson's rule. The only change is that we only integrate over the middle half of the interval $[x_i, x_{i+2}]$, so we find that the error in (10) is at most

$$\frac{M}{4!}\int_{x_{i+1/2}}^{x_{i+3/2}} |w(x)|\,dx.$$

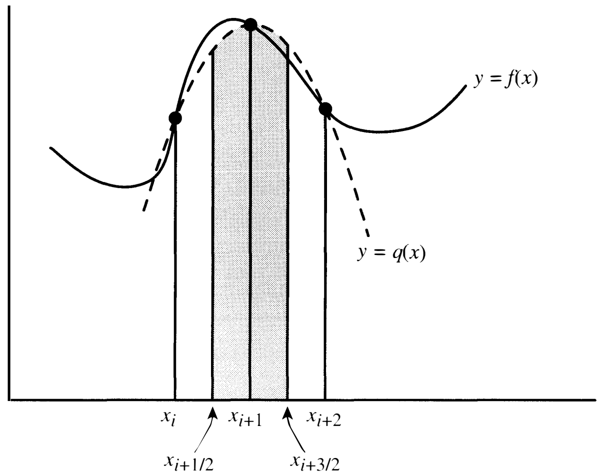


Figure 4 Approximating $\int_{x_{i+1/2}}^{x_{i+3/2}} f(x)\,dx$

Once again, we use the substitution $u = x - x_{i+1}$ to evaluate the integral of $|w(x)| = -w(x)$:

$$\int_{x_{i+1/2}}^{x_{i+3/2}} |w(x)|\,dx = \int_{x_{i+1/2}}^{x_{i+3/2}} -w(x)\,dx = \int_{-\Delta x/2}^{\Delta x/2} u^2\Delta x^2 - u^4\,du = \frac{17\Delta x^5}{240}.$$

Thus, the magnitude of the error in (10) is at most

$$\frac{M}{4!}\cdot\frac{17\Delta x^5}{240} = \frac{17\Delta x^5}{5760}M. \tag{11}$$

Notice that (10) approximates the integral of f on an interval only half as wide as the interval in the basic Simpson's rule approximation (2), but the bound (11) for the error in (10) is just a little over one-fourth of the error bound (9) for (2). Thus, we have achieved an improvement of almost a factor of two in our error bound.

Summing the approximations (10) for all integers i, $0 \le i \le n-2$, we get the approximation

$$\int_{x_{1/2}}^{x_{n-1/2}} f(x)\,dx \approx \Delta x\left[\frac{1}{24}f(x_0) + \frac{23}{24}f(x_1) + f(x_2) + f(x_3) + \cdots + f(x_{n-2}) + \frac{23}{24}f(x_{n-1}) + \frac{1}{24}f(x_n)\right]. \tag{12}$$

Since (12) is the sum of $n-1$ approximations, each of which has error bounded by (11), the magnitude of the error in (12) is at most

$$\frac{17(n-1)\Delta x^5}{5760}M. \tag{13}$$

Notice that, once again, we have a symmetrized rule. In the derivation of (12), even- and odd-numbered sample points were treated the same way, with each of the points $x_1, x_2, \ldots, x_{n-1}$ being used as the midpoint of a quadratic approximation. As a result, all coefficients in (12) except for the first two and the last two are equal to 1. However, we are not quite done, because (12) misses the intervals $[x_0, x_{1/2}]$ and $[x_{n-1/2}, x_n]$.

We will consider several ways of modifying (12) to get an approximation for $\int_a^b f(x)\,dx$. The first is simply to modify our choice of sample points so that (12) will become an approximation for $\int_a^b f(x)\,dx$. To do this, we let $\Delta x = (b-a)/(n-1)$, $x_0 = a - \Delta x/2$, and $x_i = x_0 + i\Delta x$ for $1 \leq i \leq n$. Then it is easy to verify that $a = (x_0 + x_1)/2$ and $b = (x_{n-1} + x_n)/2$, so using the new sample points in (12) instead of the old ones we get

$$\int_a^b f(x)\,dx \approx \Delta x \left[\frac{1}{24} f(x_0) + \frac{23}{24} f(x_1) + f(x_2) + f(x_3) + \cdots \right.$$
$$\left. + f(x_{n-2}) + \frac{23}{24} f(x_{n-1}) + \frac{1}{24} f(x_n) \right].$$

We will refer to this as the *central Simpson's rule*, since it is based on using only the central half of each basic Simpson's rule approximation. By (13), the magnitude of the error in this approximation is at most

$$\frac{17(n-1)\Delta x^5}{5760} M = \frac{17(b-a)^5}{5760(n-1)^4} M = \frac{17}{32} \left(\frac{n}{n-1} \right)^4 \cdot \frac{(b-a)^5}{180n^4} M, \qquad (14)$$

which is smaller than the error bound (5) for Simpson's rule when $n > 6$. As $n \to \infty$, the ratio of (14) to (5) approaches $17/32 = 0.53125$, so for large n we have cut our error almost in half.

In many situations, the central Simpson's rule would be a significant improvement over Simpson's rule. However, the central Simpson's rule also has a number of disadvantages. It requires the evaluation of f at points outside the interval $[a, b]$, so it cannot be used if f is undefined outside that interval. Furthermore, to justify the error bound (14) we need to know that f is continuous on $[x_0, x_n] = [a - \Delta x/2, b + \Delta x/2]$, and that $f^{(4)}(x)$ is defined and $|f^{(4)}(x)| \leq M$ for $x \in (x_0, x_n)$. If, for example, $f^{(4)}(x)$ grows very quickly just outside the interval $[a, b]$, then the value of M in (14) may be larger than the value of M in (5), and therefore our error bound for the central Simpson's rule may be larger than the error bound for Simpson's rule. It is therefore of interest to investigate other symmetrized versions of Simpson's rule that do not have these disadvantages. And it will turn out that this investigation will lead us to even greater improvements in accuracy.

A natural way to turn (12) into an approximation for $\int_a^b f(x)\,dx$ would be to incorporate the missed intervals $[x_0, x_{1/2}]$ and $[x_{n-1/2}, x_n]$ into our first and last applications of (10), which approximate the integrals of f on the intervals $[x_{1/2}, x_{3/2}]$ and $[x_{n-3/2}, x_{n-1/2}]$. Thus, after finding the unique quadratic polynomial q that agrees with f at the points x_0, x_1, and x_2, we would approximate the integral of f over the interval $[x_0, x_{3/2}]$ with the integral of q over the same interval. We would similarly modify our last application of (10) so that it would approximate the integral of f over the interval $[x_{n-3/2}, x_n]$, and the rest of the interval $[a, b]$ would be covered by the other applications of (10), which would remain unchanged. We will call the resulting rule the *expanded central Simpson's rule*.

Unfortunately, computing the formula for the expanded central Simpson's rule shows that it is exactly the same as the Peters-Maley rule, which, as we have already observed, has a larger error bound than Simpson's rule. Apparently, the error in the first and last intervals of the expanded central Simpson's rule can be much larger than the error in all other intervals combined, thus canceling out everything we have gained.

Intuitively, it seems that the reason the errors in the first and last intervals are so large is that these intervals are unbalanced relative to the quadratic approximation being used. For example, the integral of f over the interval $[x_0, x_{3/2}]$ is approximated using a quadratic that agrees with f at x_0, x_1, and x_2, so the interval includes more of the left half of the quadratic approximation than the right half. This lack of balance ruins the partial cancellation of errors from the two halves of the quadratic approximation that often occurs in each step of Simpson's rule. This suggests that we might be able to reduce the error by shrinking the first and last intervals in our original partition, thus making the first and last approximations more balanced.

To implement this suggestion, we will modify our partition so that $[x_0, x_1]$ and $[x_{n-1}, x_n]$ have width $r\Delta x$, for some constant r between 0 and 1, and all other intervals in the partition have width Δx. As before, we approximate $\int_{x_0}^{x_{3/2}} f(x)\,dx$ with $\int_{x_0}^{x_{3/2}} q(x)\,dx$, where q is the unique quadratic polynomial such that $q(x_0) = f(x_0)$, $q(x_1) = f(x_1)$, and $q(x_2) = f(x_2)$. This leads to the approximation

$$\int_{x_0}^{x_{3/2}} f(x)\,dx \approx \Delta x \left[\frac{(2r+1)(2r^2+2r-1)}{12r(r+1)} f(x_0) + \frac{(2r+1)^2(r+2)}{24r} f(x_1) + \frac{(2r+1)^2(1-r)}{24(r+1)} f(x_2) \right]. \tag{15}$$

We approximate the integral of f over the interval $[x_{n-3/2}, x_n]$ in a similar way, and add these approximations to the sum of the approximations (10) for $1 \le i \le n-3$. We will call the resulting approximation the *r-expanded central Simpson's rule*. (Note that the case $r = 1$ is just the expanded central Simpson's rule.) All that remains is to choose the value of r.

We want to choose r so as to minimize the error in the r-expanded rule, so we need an estimate of this error. The estimate we will use is based on Peano's theorem [**2**, pp. 285–287]. According to Peano's theorem, if $f^{(4)}(x)$ is continuous on $[a, b]$, then the error in any of the approximations we are considering is given by the formula

$$\int_a^b f^{(4)}(x)K(x)\,dx,$$

where $K(x)$ is a function called the *Peano kernel* for the approximation. Different approximations have different Peano kernels, and therefore different errors. If we assume, as usual, that $|f^{(4)}(x)| \le M$ for all $x \in (a, b)$, then the magnitude of the error is at most

$$\left| \int_a^b f^{(4)}(x)K(x)\,dx \right| \le \int_a^b \left| f^{(4)}(x)K(x) \right| dx \le M \int_a^b |K(x)|\,dx.$$

Thus, to minimize this error bound we should choose an approximation for which $\int_a^b |K(x)|\,dx$ is as small as possible.

Unfortunately, the formula for the Peano kernel is somewhat complicated, and the analysis of the Peano kernels of the r-expanded rules for different values of r is rather involved. Here, we will simply report the results of this analysis.

FIGURE 5 shows the Peano kernels for three approximations, with $n = 10$. It is clear from the figure that $\int_a^b |K(x)|\,dx$ is larger for the 1-expanded rule than for Simpson's rule. This justifies our earlier claim that the 1-expanded rule is less accurate than Simpson's rule. For large n, the optimal value of r in the r-expanded rule turns out to be $r = (\sqrt{12 + \sqrt{91}} - 2)/4 \approx 0.660264$, and it is also clear from FIGURE 5 that $\int_a^b |K(x)|\,dx$ will be significantly smaller for this rule than for Simpson's rule.

Although the optimal value of r is rather complicated, it happens to be very close to $2/3$. This suggests that $r = 2/3$ would be a good choice; the gain in simplicity from

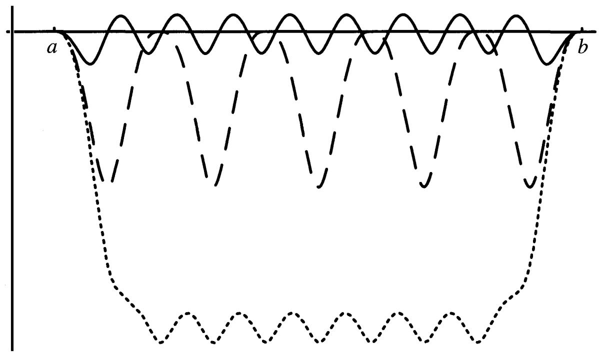


Figure 5 Peano kernels, with $n = 10$, for Simpson's rule (— — —) and the r-expanded central Simpson's rules with $r = 1$ (⋯⋯) and $r = (\sqrt{12 + \sqrt{91}} - 2)/4 \approx 0.660264$ (——)

using this value of r rather than the optimal value seems worth the resulting slight loss of accuracy.

To approximate $\int_a^b f(x)\,dx$ using the 2/3-expanded central Simpson's rule, we proceed as follows: Since the first and last intervals will have width $2\Delta x/3$ rather than Δx, we must have $b - a = (n - 2/3)\Delta x$. We therefore let $\Delta x = (b - a)/(n - 2/3)$, $x_0 = a$, $x_1 = x_0 + 2\Delta x/3$, $x_i = x_1 + (i - 1)\Delta x$ for $2 \le i \le n - 1$, and $x_n = x_{n-1} + 2\Delta x/3 = b$. To approximate the integral of f on the interval $[x_0, x_{3/2}]$ we use (15) with $r = 2/3$, which leads to the approximation

$$\int_{x_0}^{x_{3/2}} f(x)\,dx \approx \Delta x \left[\frac{77}{360} f(x_0) + \frac{49}{54} f(x_1) + \frac{49}{1080} f(x_2)\right].$$

We use a similar formula to approximate the integral of f on the interval $[x_{n-3/2}, x_n]$, and the rest of the interval $[a, b]$ is covered by the approximations (10) for $1 \le i \le n - 3$. Summing all of these approximations leads to the following formula for the 2/3-expanded rule:

$$\int_a^b f(x)\,dx \approx \Delta x \left[\frac{77}{360} f(x_0) + \frac{205}{216} f(x_1) + \frac{271}{270} f(x_2) + f(x_3) + f(x_4) + \cdots + f(x_{n-3}) + \frac{271}{270} f(x_{n-2}) + \frac{205}{216} f(x_{n-1}) + \frac{77}{360} f(x_n)\right].$$

Of course, this is another symmetrized rule, with all coefficients except the first three and the last three equal to 1. We can find a bound for the error by integrating the absolute value of the Peano kernel and applying Peano's theorem. This calculation shows that the magnitude of the error in the 2/3-expanded rule is at most

$$\left(\alpha - \frac{\beta}{n}\right)\left(\frac{n}{n - 2/3}\right)^5 \cdot \frac{(b - a)^5}{180 n^4} M, \tag{16}$$

where M is, as usual, an upper bound on $|f^{(4)}(x)|$, α is given by the formula

$$\alpha = \frac{171 + 2\sqrt{81 - 12\sqrt{30}} + 36\sqrt{270 - 40\sqrt{30}}}{2916} \approx 0.149411,$$

and β is a constant whose formula is too complicated to print here, but whose numerical value is approximately 0.0309389. For large n, this error bound is about 0.149411 times the error bound for Simpson's rule, an improvement by a factor of almost 7.

The error bounds for the r-expanded rules with other values of r have the same form, but with $n - 2/3$ replaced by $n - 2 + 2r$, and with different values for the constants α and β. The optimal value of r given earlier is optimal in the sense that it leads to the smallest value of α. This smallest value is $75/512 \approx 0.146484$, so the increase in error that results from using $r = 2/3$ rather than the optimal value of r is fairly small.

In addition to greater simplicity, the use of $r = 2/3$ has another modest advantage: Suppose we use the 2/3-expanded rule to compute an approximation of $\int_a^b f(x)\,dx$, and then we decide that we want to compute a more accurate approximation by increasing the value of n. If we increase n to $4n - 2$, then it is easy to verify that the new value of Δx will be exactly 1/4 of the old one, and all the old sample points will be among the new sample points. Thus, we can reuse all of our function evaluations from the first use of the 2/3-expanded rule.

The 2/3-expanded rule avoids the problems of the central Simpson's rule, because it does not involve the use of sample points outside of the interval $[a, b]$. However, it does have one limitation: It cannot be used if the only information we have about f is a table of values at evenly spaced sample points. Thus, it would be interesting to know if it is possible to formulate a rule based on (12) that uses exactly the same sample points as Simpson's rule.

The difficulty in formulating such a rule is, as usual, that we must find a way to approximate the integrals of f over the troublesome intervals $[x_0, x_{1/2}]$ and $[x_{n-1/2}, x_n]$. The simplest way to deal with these intervals while guaranteeing that the final approximation will be exact for cubic polynomials is to use a cubic approximation for each of these intervals. For example, we might approximate the integrals of f over these intervals with the integrals of cubic polynomials that agree with f at the first four and last four sample points. We will not pursue this approach here, since it violates the spirit of this paper, which is to use only quadratic approximations. However, we note that the resulting formula would be exactly the same as the formula $Q_n^{4,2}$ derived by Peter Köhler [**5**]. The error bound for large n is approximately 17/32 times the error bound for Simpson's rule.

Examples

FIGURE 6 shows the ratios of our bounds on the errors in the central Simpson's rule and the 2/3-expanded rule (formulas (14) and (16)) to the error bound for Simpson's rule (formula (5)), as functions of n. In all of these error bounds, the coefficients are the best possible. For Simpson's rule and the central Simpson's rule, this can be seen by considering $f(x) = x^4$. This function has a constant fourth derivative, $f^{(4)}(x) = 4! = 24$, so we can use $M = 24$ in our error bounds. But then M is not just an upper bound on the magnitude of the quantity $f^{(4)}(c_x)$ that appears in the lemma, it is actually equal to that quantity. It follows that our error bound calculations actually give the exact magnitudes of the errors involved for this function.

However, the magnitude of the error in the 2/3-expanded rule for the function $f(x) = x^4$ is smaller than the error bound (16) for that rule. The reason is that according to Peano's theorem, the error in the 2/3-expanded rule in this case is

$$\left|\int_a^b f^{(4)}(x)K(x)\,dx\right| = M\left|\int_a^b K(x)\,dx\right|,$$

and this is smaller than our bound $M\int_a^b |K(x)|\,dx$, since the Peano kernel for the 2/3-expanded rule is sometimes positive and sometimes negative. However, it can be

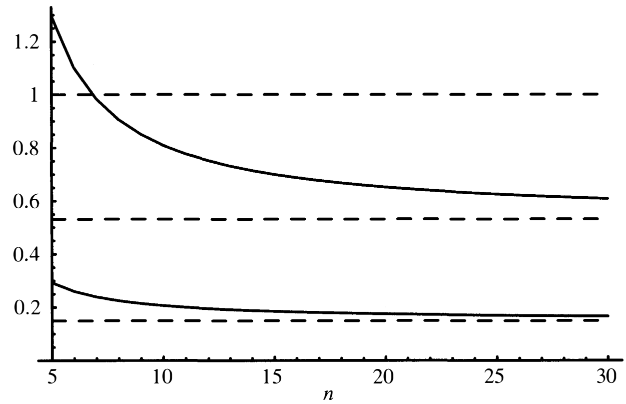


Figure 6 Ratios of error bounds for the central Simpson's rule (upper curve) and the 2/3-expanded rule (lower curve) to the error bound for Simpson's rule, as functions of n, for $5 \le n \le 30$ (The dashed lines are at 1, 17/32, and 0.149441.)

shown, by using an example that is more complicated than x^4, that our error bound for the 2/3-expanded rule is the best possible.

These observations are confirmed by TABLE 1, which shows the values, errors, and error bounds for the approximations of the integral $\int_0^1 x^4\,dx = 0.2$ by Simpson's rule, the central Simpson's rule, and the 2/3-expanded rule, using $n = 10$, $n = 20$, and $n = 100$. All of the errors are negative, indicating that our approximations are larger than the exact value of the integral. As expected, the magnitudes of all of the errors for Simpson's rule and the central Simpson's rule are exactly equal to the bounds given by (5) and (14), but the magnitudes of the errors for the 2/3-expanded rule are smaller than the bounds given by (16).

TABLE 1: Approximation of $\int_0^1 x^4\,dx = 0.2$

		Simpson's rule	Central rule	2/3 rule
$n = 10$	Value	0.20001333	0.20001080	0.20000144
	Error	-1.333×10^{-5}	-1.080×10^{-5}	-1.437×10^{-6}
	Error Bound	1.333×10^{-5}	1.080×10^{-5}	2.755×10^{-6}
$n = 20$	Value	0.200000833	0.200000544	0.200000067
	Error	-8.333×10^{-7}	-5.435×10^{-7}	-6.663×10^{-8}
	Error Bound	8.333×10^{-7}	5.435×10^{-7}	1.460×10^{-7}
$n = 100$	Value	0.20000000133	0.20000000074	0.20000000008
	Error	-1.333×10^{-9}	-7.374×10^{-10}	-8.329×10^{-11}
	Error Bound	1.333×10^{-9}	7.374×10^{-10}	2.056×10^{-10}

Next we consider the integral $\int_0^{20} \sin x\,dx = 1 - \cos 20 \approx 0.5919179382$. FIGURE 7 shows the approximations of this integral using Simpson's rule, the central Simpson's rule, and the 2/3-expanded rule, with $n = 10$ in all cases. In each graph, the dashed lines are the quadratic approximations to $y = \sin x$, the black dots are the sample points, and the shaded region is the region whose area is being used to approximate the integral.

A striking problem with Simpson's rule is evident in FIGURE 7(a) at the fifth sample point, which occurs near the second local maximum of the curve. This point is in the

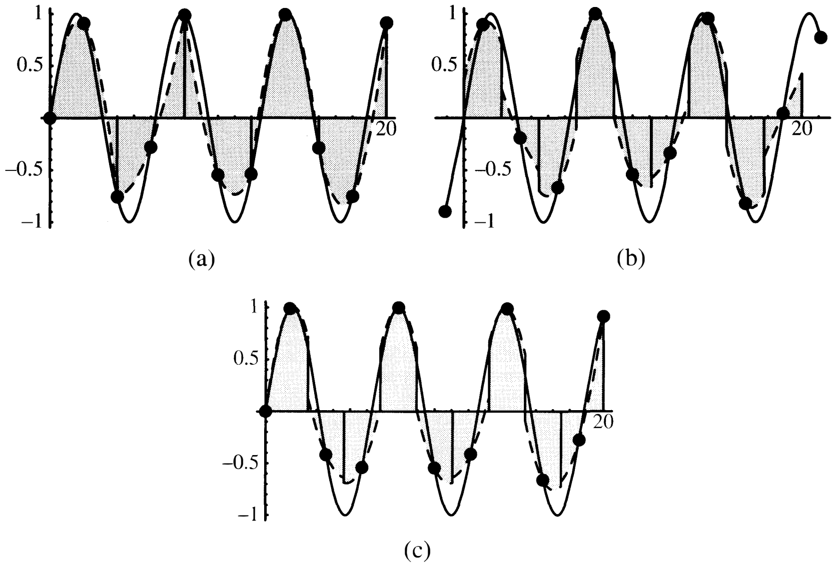


Figure 7 Approximation of $\int_0^{20} \sin x\, dx$ using (a) Simpson's rule, (b) the central Simpson's rule, and (c) the 2/3-expanded rule, with $n = 10$

middle of an interval on which the curve is concave down, but the Simpson's rule approximation does not detect this fact, because this sample point is the dividing point between two intervals on which $\sin x$ is approximated by quadratic polynomials, not a midpoint of such an interval. This problem does not occur in FIGURES 7(b) and 7(c).

TABLE 2 shows the values of our various approximations for this integral, and their errors and error bounds, for $n = 10$, $n = 20$, and $n = 100$. As expected, the central Simpson's rule is more accurate than Simpson's rule, and the 2/3-expanded rule is the most accurate of the three.

TABLE 2: Approximation of $\int_0^{20} \sin x\, dx \approx 0.5919179382$

		Simpson's rule	Central rule	2/3 rule
$n = 10$	Value	0.68735	0.63563	0.61913
	Error	−0.09543	−0.04371	−0.02721
	Error Bound	1.778	1.439	0.3673
$n = 20$	Value	0.595644	0.594072	0.592909
	Error	−0.003726	−0.002154	−0.000991
	Error Bound	0.11111	0.07247	0.01946
$n = 100$	Value	0.591923225	0.591920848	0.591918449
	Error	-5.287×10^{-6}	-2.910×10^{-6}	-5.109×10^{-7}
	Error Bound	1.778×10^{-4}	9.832×10^{-5}	2.741×10^{-5}

Finally, we provide in TABLE 3 some calculations for the integral

$$\int_0^1 \frac{4}{1+x^2}\, dx = \pi.$$

Surprisingly, although the central Simpson's rule does much better than Simpson's rule on this integral, the 2/3-expanded rule does much worse.

TABLE 3: Approximation of $\int_0^1 \frac{4}{1+x^2}\,dx = \pi$

		Simpson's rule	Central rule	2/3 rule
$n = 10$	Error	3.965×10^{-8}	1.166×10^{-9}	-6.271×10^{-7}
	Error Bound	5.333×10^{-5}	4.318×10^{-5}	1.102×10^{-5}
$n = 20$	Error	6.200×10^{-10}	1.318×10^{-11}	-1.793×10^{-8}
	Error Bound	3.333×10^{-6}	2.174×10^{-6}	5.839×10^{-7}
$n = 100$	Error	3.968×10^{-14}	6.586×10^{-16}	-5.187×10^{-12}
	Error Bound	5.333×10^{-9}	2.950×10^{-9}	8.223×10^{-10}

All of the tables and graphs for the examples in this section were created using *Mathematica.* Readers who want to try more examples themselves can find a *Mathematica* notebook that computes all of our integration rules at `www.maa.org/pubs/mathmag.html`.

Conclusion and extensions

We have found several ways to improve on Simpson's rule while treating the even- and odd-numbered sample points symmetrically. The key idea behind these improvements is to fit quadratic polynomials to f at triples of successive sample points spanning intervals of width $2\Delta x$, as in the usual Simpson's rule, but then to integrate these polynomials over intervals of width only Δx.

This idea can also be applied to the other Newton-Cotes numerical integration formulas, which are all based on approximating f with polynomials. As an example, we briefly discuss the application of this idea to Boole's rule [**2**, p. 78]. To approximate $\int_a^b f(x)\,dx$ by Boole's rule, we begin by dividing the interval $[a, b]$ into n subintervals of width $\Delta x = (b-a)/n$, where n is a multiple of 4. Then for each i that is a multiple of 4, $0 \leq i \leq n-4$, we approximate f on the interval $[x_i, x_{i+4}]$ with a polynomial of degree 4 that agrees with f at all five of the sample points in this interval, and approximate the integral of f on this interval with the integral of the polynomial. Summing all of these approximations yields the approximation

$$\int_a^b f(x)\,dx \approx \Delta x\left[\frac{14}{45}f(x_0) + \frac{64}{45}f(x_1) + \frac{8}{15}f(x_2) + \frac{64}{45}f(x_3)\right.$$
$$+ \frac{28}{45}f(x_4) + \frac{64}{45}f(x_5) + \frac{8}{15}f(x_6) + \frac{64}{45}f(x_7)$$
$$\left. + \frac{28}{45}f(x_8) + \cdots + \frac{64}{45}f(x_{n-1}) + \frac{14}{45}f(x_n)\right].$$

To improve and symmetrize Boole's rule, we again find a polynomial of degree 4 agreeing with f at all sample points in the interval $[x_i, x_{i+4}]$, but we only integrate this polynomial over an interval of width Δx centered at x_{i+2}, to get an approximation of $\int_{x_{i+3/2}}^{x_{i+5/2}} f(x)\,dx$. Summing these approximations for $0 \leq i \leq n-4$ yields the approximation

$$\int_{x_{3/2}}^{x_{n-3/2}} f(x)\,dx \approx \Delta x \left[-\frac{17}{5760} f(x_0) + \frac{97}{1920} f(x_1) + \frac{1823}{1920} f(x_2) + \frac{5777}{5760} f(x_3) + f(x_4) + f(x_5) + \cdots + f(x_{n-4}) + \frac{5777}{5760} f(x_{n-3}) + \frac{1823}{1920} f(x_{n-2}) + \frac{97}{1920} f(x_{n-1}) - \frac{17}{5760} f(x_n) \right]. \qquad (17)$$

The error bound for this approximation is about $367/2048 \approx 0.179$ times the error bound for Boole's rule. To get an approximation for $\int_a^b f(x)\,dx$, we would need to modify (17) to get it to cover the entire interval $[a, b]$. We leave the details of this modification to the reader.

Acknowledgment. I would like to thank Norton Starr and several anonymous referees for helpful comments on earlier drafts of this paper.

REFERENCES

1. Tom M. Apostol, *Calculus*, Vol. II, 2nd ed., John Wiley & Sons, New York, 1969.
2. Philip J. Davis and Philip Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, Orlando, 1984.
3. J. M. De Villiers, A nodal spline interpolant for the Gregory rule of even order, *Numer. Math.* **66** (1993), 123–137.
4. Herman H. Goldstine, *A History of Numerical Analysis From The 16th Through The 19th Century*, Springer-Verlag, New York, 1977.
5. Peter Köhler, On a generalization of compound Newton-Cotes quadrature formulas, *BIT* **31** (1991), 540–544.
6. G. O. Peters and C. E. Maley, Numerical integration over any numbers of equal intervals, *Amer. Math. Monthly* **75** (1968), 741–744.
7. G. M. Phillips, Gregory's method for numerical integration, *Amer. Math. Monthly* **79** (1972), 270–274.
8. R. S. Pinkham, Simpson's rule with constant weights, *College Math. J.* **32** (2001), 91–93.
9. Kenneth J. Supowit, Understanding the extra power of the Newton-Cotes formula for even degree, this MAGAZINE **70** (1997), 292–293.

NOTES

Tiling Deficient Rectangles with Trominoes

J. MARSHALL ASH
DePaul University
Chicago, IL 60614
mash@math.depaul.edu

SOLOMON W. GOLOMB
Communication Sciences Institute
University of Southern California
Los Angeles, CA 90089-2565

A *tromino* (rhymes with domino) is a shape made up of three 1×1 squares assembled as shown.

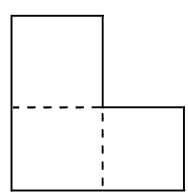


Figure 1 A tromino

We will classify a variety of nearly rectangular shapes into those that can be tiled by trominoes and those that cannot. From now on we will simply say *tiled* to mean tiled by trominoes. We will consider shapes that are integer-dimensioned rectangles with each dimension at least 2, and with one or two 1×1 squares removed. If one square is removed, call the resultant shape a *deficient rectangle*. If the removed square was a corner square, call the resulting deficient rectangle a *dog-eared rectangle*. The area of a tromino is 3, so, evidentially, only shapes whose area is a multiple of 3 can be tiled by trominoes. In this paper we will determine which deficient rectangles with area divisible by 3 are tileable and which are not; in particular, all the dog-eared ones are tileable. We will also get some partial results for the same question for rectangles with two squares removed and remaining area divisible by 3.

We especially recommend the proof of the Deficient 5×5 Lemma to the casual reader.

Trominoes were introduced by Golomb [**3**], who proved that deficient squares whose side length is a power of two can be tiled. Chu and Johnsonbaugh first extended Golomb's work to the general cases of deficient squares [**1**]. They later went on to rectangles and proved a slightly weaker version [**2**] of what we call the Deficient Rectangle Theorem. The Proposition in the last section answers a question posed by Chu and Johnsonbaugh [**1**].

Before proceeding with the business at hand, we will mention a few general facts about trominoes and about another tiling question involving tiles other than trominoes.

A tromino is a special case of a polyomino, a shape made by connecting certain numbers of 1×1 squares, each joined together with at least one other square along an edge. The polyomino of area 1, a single 1×1 tile is called a *monomino*. The polyomino of area 2 is the domino. Let $P(n)$ be the number of distinct polyominoes of area n. For example, $P(3) = 2$, since there are actually two trominoes: a straight tromino, which has the shape of a 1×3 rectangle, and the object shown in FIGURE 1, which is called a right tromino when it needs to be distinguished from the straight one, but which will be the only kind of tromino discussed in this paper. Notice that in defining P, orientation is ignored. For example, the three objects created by rotating the tromino shown in FIGURE 1 by 90°, 180°, and 270° are not counted as distinct from the original tromino. For results and open questions about the exponentially growing values of $P(n)$, see [**4**, Appendix D].

Four trominoes can be fit together to form a tromino-shaped 4-*reptile*, that is, a set in the plane that can be tiled by four congruent scaled down copies of itself. A tromino has *order* 2, which means that the minimum number of trominoes required to form a rectangle is 2, as in FIGURE 2. Finding the order of other polyominoes provides challenging problems [**4**, Chapter 8]. The entire plane can be tiled in a periodic way by any polyomino of finite order by simply repeating copies of the minimal rectangle.

Roger Penrose has given a remarkably simple aperiodic tiling of the entire plane using copies of only two unit-edged rhombi, one with acute angle 36° and the other with acute angle 72° [**8**]. Tilings are often found in Moorish architecture; some trominoes can be seen in a display case in the *Reales Alcazares*, a great Arabian style palace built during various epochs in Seville, Spain. A comprehensive and interesting book concerning tiling is *Tilings and Patterns* [**5**]. There is lots of information about tiling available on the internet; typing "tromino" into a search engine produced 577 hits. We recommend `http://www.ics.uci.edu/~eppstein/junkyard/polyomino.html` and `http://www.amherst.edu/~nstarr/`.

Elementary results for rectangles A basic tiling result that we will need identifies precisely which rectangles can be tiled. Let's start with some simple cases. First of all, a 2×3 rectangle can be tiled by trominoes.

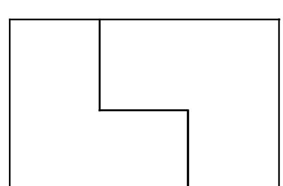


Figure 2 Tiling $R(2, 3)$

Denote a rectangle with i rows and j columns by $R(i, j)$. We will indicate decompositions into nonoverlapping subrectangles by means of an additive notation. For example, a $3i \times 2j$ rectangle can be decomposed into ij 3×2 subrectangles and we write this fact as $R(3i, 2j) = \sum_{\mu=1}^{i} \sum_{\nu=1}^{j} R(3, 2) = ijR(3, 2)$. It follows from this and the tiling in FIGURE 2 that

$$\text{any } 3i \times 2j \text{ or } 2i \times 3j \text{ rectangle can be tiled.} \quad (1)$$

From now on, any rectangle decomposed into a combination of $3i \times 2j$ subrectangles, $2i \times 3j$ subrectangles, and trominoes will be considered as successfully tiled by trominoes. Denote the 1×1 square lying in row i and column j as (i, j).

Now let's look at some rectangles that cannot be tiled. Suppose that a 3×3 square Q has been tiled. Some tromino must cover square $(3, 1)$. Here are the three possible ways that it can do that.

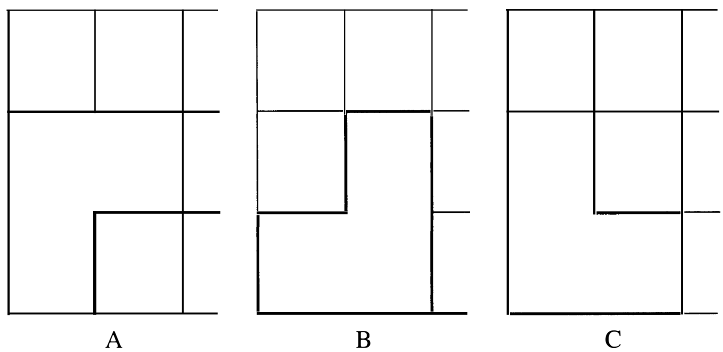


Figure 3 An impossibility proof

Orientation A is immediately ruled out, since square $(1, 1)$ cannot be tiled. But in cases B and C, the tiling must tile the leftmost 3×2 subrectangle of Q, so that the original tiling is also a tiling of the third column of Q, which is an $R(3, 1)$. This is impossible. Similarly, suppose that a 3×5 rectangle R has been tiled. This argument shows that the tiling must tile the first two columns of R, and hence also the rightmost three columns of R. This is a contradiction since we have just shown 3×3 square to be untileable. Iterating this procedure shows no $R(3, \text{odd})$ can be tiled. It turns out that there are no other untileable rectangles with area divisible by 3.

The integers m and n will always be greater than or equal to 2.

CHU-JOHNSONBAUGH THEOREM [**2**]. *An $m \times n$ rectangle can always be tiled by trominoes if 3 divides its area mn, except when one dimension is 3 and the other is odd.*

The proof of this is not hard. We suggest that the reader give it a try. Here are a few hints. Use fact (1) several times. First do the cases $R(3k, \text{even})$; then do the cases $R(6k, \text{odd})$. This leaves only the cases $R(9 + 6k, n)$, where $n \geq 5$ is an odd integer. Reduce such a case to $R(9, 5)$. Finally tile $R(9, 5)$ by trial and error. If you have trouble with the last step, leaf ahead to the top left picture in FIGURE 9.

Dog-ears An $m \times n$ *dog-eared rectangle* is an $m \times n$ rectangle with a 1×1 corner square removed. We will denote the dog-eared rectangle by $R(m, n)^-$, so that $R(m, n)^- = R(m, n) \setminus \{(1, n)\}$. Note that the area of $R(m, n)^-$ is $mn - 1$. If this rectangle is rotated 180°, a similar figure with missing lower left-hand corner is created. If it is reflected about a central vertical (resp., horizontal) axis, a similar figure with missing upper left-hand (resp., lower right-hand) corner is created. The problem of tiling the original figure is clearly equivalent to tiling any one of the other three, even though the original figure cannot be rotated into either of the last two figures.

DOG-EARED RECTANGLE THEOREM. *An $m \times n$ dog-eared rectangle can be tiled with trominoes if and only if 3 divides its area.*

To understand what this theorem means, note that if mn is congruent to 0 or 2 modulo 3, then the area of $R(m, n)^-$ is not congruent to 0 modulo 3 and so that dog-eared rectangle cannot be tiled by trominoes, since the area of any region tiled by trominoes must be an integral multiple of 3. So the only $m \times n$ dog-eared rectangles that

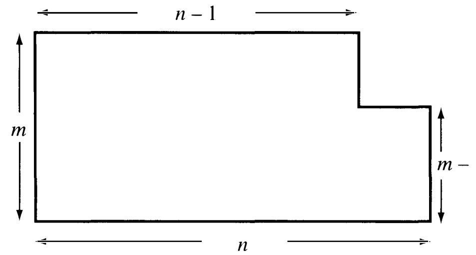


Figure 4 A dog-eared rectangle

could possibly be tiled by trominoes are those for which mn is congruent to 1. In other words, the only $m \times n$ dog-eared rectangles that could possibly be tiled by trominoes are those for which $m \equiv n \equiv 1 \pmod 3$ or $m \equiv n \equiv 2 \pmod 3$, and, indeed, all those dog-eared rectangles can be tiled.

We start with a family of special cases of the Dog-eared Rectangle Theorem, the dog-eared dyadic squares, $R(2^k, 2^k)^-$. This is a special case of a well-known and beautiful example of mathematical induction [**3**], [**4**, page 4], [**6**, page 45], [**9**, problem 2.3.38]. If $k = 1$, note that $R(2, 2)^-$ is itself a tromino. If $k = 2$, see FIGURE 5 for a covering of $R(4, 4)^-$. In FIGURE 5, $R(4, 4)^-$ was tiled by dividing it into 4 quadrants. The upper right quadrant was an $R(2, 2)^-$ while the other three quadrants were all congruent to $R(2, 2)$. Then the black tromino covering squares $(2, 2)$, $(2, 3)$, and $(3, 3)$ was placed at the center. This reduced the lower left quadrant to an $R(2, 2)^-$ and the remaining two quadrants to rotations of $R(2, 2)^-$. In short, the tiling of $R(4, 4)^-$ was reduced to the tiling of four copies of $R(2, 2)^-$. The reader should next do $k = 3$, by dividing $R(8, 8)^-$ into 4 quadrants, and then covering the central squares $(4, 4)$, $(4, 5)$, and $(5, 5)$ with a tromino. This reduces the tiling of $R(8, 8)^-$ to the tiling of four copies of $R(4, 4)^-$. The general inductive proof should now be clear.

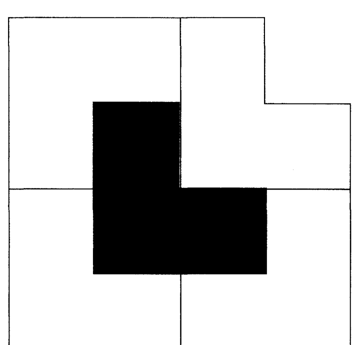


Figure 5 Tiling $R(4, 4)^-$

Proof of the Dog-eared Rectangle Theorem: Let $m \leq n$. As mentioned above, the necessary condition that 3 divide the area of $R^- = R(m, n)^-$ splits into the cases $m \equiv n \equiv 1 \pmod 3$ and $m \equiv n \equiv 2 \pmod 3$. It is not hard to see that the $m \times n$ dog-eared rectangle is congruent to the $n \times m$ one. So, in the former case, we have either $R(4, 3k + 4)^-$ with $k \geq 0$, $R(7, 6k + 7)^-$ with $k \geq 0$, $R(7, 6k + 4)^-$ with $k \geq 1$, or $R(3j + 4, 3k + 4)^-$ with $j \geq 2$ and $k \geq 2$. There correspond these four decompositions:

$$R(4,3k+4)^- = R(4,3k) + R(4,4)^-, k \geq 0$$
$$R(7,6k+7)^- = R(7,6k) + R(7,7)^-, k \geq 0$$
$$R(7,6k+4)^- = R(7,6k) + R(4,3) + R(4,4)^-, k \geq 1 \text{ and}$$
$$R(3j+4,3k+4)^- = R(3j,3k+4) + R(4,3k) + R(4,4)^-, j,k \geq 2.$$

For the algebraically inclined reader, these decompositions need no further explanation. However, the geometrically inclined reader should draw pictures to visualize them. (All the similar decompositions appearing below have straightforward geometrical interpretations.) Here, in the first three cases, a large rectangle was stripped from the left side of the figure. In the fourth case, first a large rectangle was removed from the bottom, and then another from the left side of what remained. All the full rectangles are tileable by the Chu-Johnsonbaugh Theorem, $R(4,4)^-$ is tiled as in FIGURE 5, and the tiling of $R(7,7)^-$ appears in Chu and Johnsonbaugh [**1**].

In the latter case, we must tile $R^- = R(3j+2, 3k+2)^-$ where $0 \leq j \leq k$. If $j \neq 1$, we have

$$R^- = R(3j,3k) + R(3j,2) + R(2,3k) + R(2,2)^-.$$

The first three terms are tiled by the Chu-Johnsonbaugh Theorem, while the last term actually *is* a tromino. Let $j = 1$. Either k is odd, $3k + 2 = 6\ell + 5$; or else k is even, $3k + 2 = 6\ell + 8$. Correspondingly, either $R^- = R(5, 6\ell + 5)^- = R(5, 6\ell) + R(5,5)^-$ where the first term is tiled with the Chu-Johnsonbaugh Theorem and $R(5,5)^-$ is tiled as in FIGURE 6, or else $R^- = R(5, 6\ell + 8)^- = R(5, 6\ell) + R(5,8)^-$ where again the first term is tiled with the Chu-Johnsonbaugh Theorem and we also have $R(5,8)^- = R(5,6) + R(3,2) + R(2,2)^-$, the first two terms being tiled by the Chu-Johnsonbaugh Theorem, while the last term is a tromino. ■

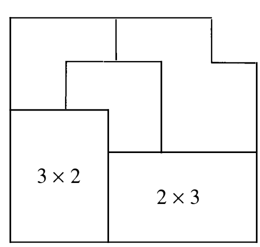


Figure 6 Tiling $R(5,5)^-$

Here is an application of the Chu-Johnsonbaugh and Dog-eared Rectangle Theorems. Consider the practical question of tiling as much as possible of any $m \times n$ rectangle, where m and n both exceed 3. There are 3 cases depending on the value of mn modulo 3. If $mn \equiv 0$, tile the entire rectangle with the Chu-Johnsonbaugh Theorem. If $mn \equiv 1$, remove a single corner square and then use the Dog-eared Rectangle Theorem to tile the rest of the rectangle. If $mn \equiv 2$, we must remove 2 squares. It turns out that if $mn \equiv 2$ and if a corner square and a boundary square adjacent to it are both removed, what remains can always be tiled. This can be proved by methods very similar to those used to prove the other two theorems. We will leave its proof as an exercise.

Deficiency Call a rectangle with one 1×1 square missing a *deficient rectangle*. Thus to the rectangle $R(m,n)$ correspond mn deficient rectangles, each being formed by removing one square from $R(m,n)$; exactly 4 of these are congruent to $R(m,n)^-$. The question of whether a deficient rectangle can be tiled with trominoes is clearly equivalent to the question of whether the full rectangle consisting of the disjoint union of the deficient rectangle and the 1×1 square can be tiled by a set of trominoes and a single monomino, with the monomino covering the missing square.

Say that a 1×1 square is *good* if its removal from a full $m \times n$ rectangle produces a deficient rectangle that can be tiled. We will now enumerate some $m \times n$ deficient rectangles that cannot be tiled, even though 3 divides $mn - 1$. This enumeration will be very precise in the sense that for each m and n the location of the bad squares will be specified. In the very interesting case when $m = n = 5$, the following lemma produces 16 bad squares.

DEFICIENT 5×5 LEMMA. *If the square (i, j) is removed from the 5×5 rectangles where either i or j is even, then the resulting shape is not tileable.*

Proof. Form a kind of checkerboard design by marking each of the nine squares

$$\left\{ \begin{array}{lll} (1,1), & (1,3), & (1,5), \\ (3,1), & (3,3), & (3,5), \\ (5,1), & (5,3), & (5,5), \end{array} \right\}$$

and assume that one of the 16 unmarked squares has been removed from $R(5,5)$ to form R^-. Then a proposed tiling of R^- must contain one tromino for each of the 9 marked squares, so that tiling must have area at least $9 \cdot 3 = 27$, which is absurd since the area of R^- is 24. Thus all 16 of the unmarked squares are bad. ■

Next we note that bad squares can also occur when $(m,n) = (2, 5 + 3k)$, $k = 0, 1, 2, \ldots$. Here some bad squares are those of the form $(x, 3j)$, $j = 1, 2, \ldots, k+1$, $x = 1$ or 2. By symmetry we may assume that $x = 1$. To show that $R(2, 5+3k) \setminus \{(1, 3j)\}$ cannot be tiled, assume the opposite and let T be the tromino covering the square $(2, 3j)$. Then to the left of $T + \{(1, 3j)\}$ lies either the rectangle $R(2, 3j - 1)$ or the rectangle $R(2, 3j - 2)$, neither of which has area divisible by 3.

Finally, the square $(3, 2)$ is bad in the $5 \times (5 + 3k)$ case, that is, $R(5, n) \setminus \{(3, 2)\}$ cannot be tiled. For if $(3, 2)$ were good, some tromino T would have to cover the square $(3, 1)$. If T lay above $(3, 2)$ the square $(1, 1)$ could not be reached, otherwise the square $(5, 1)$ could not be reached. Symmetrically, $(3, 4 + 3k)$ is also bad in this case.

DEFICIENT RECTANGLE THEOREM (COMPARE [**2**]). *An $m \times n$ deficient rectangle, $2 \le m \le n$, $3 | mn - 1$, has a tiling, regardless of the position of the missing square, if and only if (a) neither side has length 2 unless both of them do, and (b) $m \ne 5$. Furthermore, in all the exceptional cases the only bad squares are those enumerated in the preceding discussion.*

Proof. For this proof only, we change notation slightly and let $R(m,n)^-$ denote any $m \times n$ rectangle of deficiency 1. The "outlier" $R(2,2)^-$ is tiled with one tromino. First assume that $m \ge 4$, $m \ne 5$, and $3 \nmid m$. The method of proof is to proceed inductively after treating the cases $m = 4, 7, 8, 10$, and 11 individually. If $m \ge 13$, then $m - 6 > 6$ so that we may slice a full rectangle of height 6 off of either the top or the bottom of $R(m,n)^-$, that is, $R(m,n)^- = R(m-6, n)^- + R(6, n)$. Since the last term is tileable by the Chu-Johnsonbaugh Theorem, this first reduces the cases $m \in [13, 17]$ to the cases $m \in [7, 11]$, then the cases $m \in [19, 23]$ to the cases $m \in [13, 17]$, and so on.

If $m = 4$, write $R(4, 3k+1)^- = (k-1)R(4,3) + R(4,4)^-$. Apply the Chu-Johnsonbaugh Theorem to the first $k - 1$ terms. For the last term, observe that in [**3**], Golomb showed that all $2^k \times 2^k$ deficient squares can be tiled. (Its proof is an induction argument almost identical to the one used above to tile the $2^k \times 2^k$ dog-eared squares.) If $m = 7$, we may write $R(7, n)^- = R(3, n) + R(4, n)^-$ and thus reduce the $m = 7$ case to the $m = 4$ case when n is even; while if n is odd, 6 divides $n - 7$ and $R(7, n)^- = ((n - 7)/6)R(7, 6) + R(7, 7)^-$ is tiled using the Chu-Johnsonbaugh Theorem and reference [**2**]. If $m = 10$; then $R(10, n)^- = R(7, n)^- + R(3, n)$ so that the Chu-Johnsonbaugh Theorem provides a reduction to the $m = 7$ case if n is even, while $R(10, n)^- = R(10, n - 3)^- + R(10, 3)$ reduces the odd n case to the even n case. If $m = 8$, $R(8, n)^- = R(8, 8 + 3k)^- = kR(8, 3) + R(8, 8)^-$ is tiled by the Chu-Johnsonbaugh Theorem and reference [**3**]. Finally if $m = 11$, n must be congruent to either 8 or 11 modulo 6. If $n = 8 + 6k$, $R(11, 8 + 6k)^- = kR(11, 6) + R(11, 8)^-$ with the first terms tiled by the Chu-Johnsonbaugh Theorem and the last term tiled by the $m = 8$ case since $R(11, 8)^- = R(8, 11)^-$, while if $n = 11 + 6k$, $R(11, 11 + 6k)^- = kR(11, 6) + R(11, 11)^-$; the first terms are tiled by the Chu-Johnsonbaugh Theorem and the tiling of the last term can be found in reference [**1**].

In view of the treatment of all the bad cases before the statement of this theorem, it remains only to analyze the exceptional good cases. Since 3 divides $mn - 1$, if $m = 2$, we must have $n = 2 + 3k, k = 0, 1, \ldots,$ while if $m = 5$, we must have $n = 5 + 3k, k = 0, 1, 2, \ldots$. Also notice that the $(m, n) = (2, 2)$ case is *not* an exception. We'll start with the 5×5 good cases. The tiling in FIGURE 6 above shows that $(1, 5)$ is good, while these two tilings show $(3, 5)$ and $(3, 3)$ to be good. Symmetry considerations show that the remaining six marked tiles are also good. Thus all nine marked tiles are good.

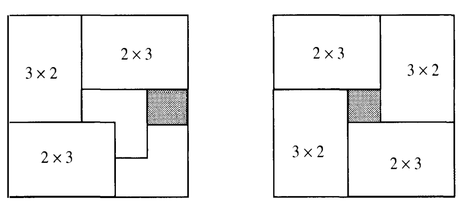


Figure 7 Tiling deficient 5×5 rectangles

Next, if $(m, n) = (2, 5 + 3k), k = 0, 1, \ldots,$ we determine to be good all the squares of the form $(x, 3j + 1)$ or $(x, 3j + 2), j = 0, 1, \ldots, k + 1$, where $x = 1$ or 2. In fact, we may write any of these as $(k + 1)R(2, 3) + R(2, 2)^-$, apply (1) to each of the first $k + 1$ terms, and use one more tromino to cover $R(2, 2)^-$.

It remains to treat the deficient rectangles $R(5, 8 + 3k)^-$, where $k \geq 0$ and the removed square is neither $(3, 2)$ nor $(3, 7 + 3k)$. Assume that all the cases $R(5, 8)^-$ and $R(5, 11)^-$ have been done and that any square removed from now on is not $(3, 2)$. Let the square (i, j) be removed from $R(5, 14)$. Symmetry allows the assumption $j \leq 7$. If $(i, j) \neq (3, 7)$, then the decomposition of the resulting $R(5, 14)^-$ into an $R(5, 8)^-$ on the left and an $R(5, 6)$ on the right allows a tiling, while $R(5, 14) \setminus \{(3, 7)\}$ is tiled by decomposing it into an $R(5, 6)$ on the left and an $R(5, 8)^-$ on the right. The cases of $R(5, n), n \geq 17$ will be treated inductively. Symmetry allows us to consider only $R(5, n) \setminus \{(i, j)\}$ where $j \leq n/2 < n - 8$ and where all but 2 tiles of $R(5, n - 6)$ are good. Now decompose into $R(5, n - 6)^-$ on the left and $R(5, 6)$ on the right. Since $j \neq (n - 6) - 1$ the first term may be tiled, while the second is tiled with the Chu-Johnsonbaugh Theorem.

The cases $R(5, 8)$. By symmetry we may assume $i \geq 3$ and $j \leq 4$. Since $(3, 2)$ is bad, we have 11 cases to show good. If $i \geq 4$ and $j \in \{1, 2, 4\}$, then (i, j) is a good square of $R(2, 8)$, so the decomposition of $R(5, 8)^-$ into a full upper rectangle $R(3, 8)$ and a lower $R(2, 8)^-$ works in all six of these cases. There remain the five cases $(i, j) = (3, 1), (5, 3), (4, 3), (3, 3)$, and $(3, 4)$. These are done in ad hoc fashion in FIGURE 8.

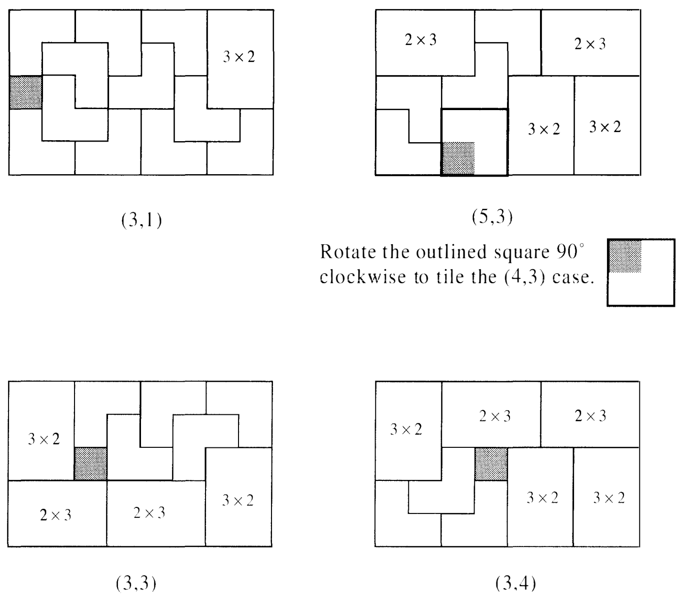


Figure 8 Deficient 5×9 tilings

The cases $R(5, 11)$. By symmetry we may assume $i \geq 3$ and $j \leq 6$. Since $(3, 2)$ is bad, we have 17 cases to show good. If i and j are both odd, (i, j) is a good square of $R(5, 5)$, so the decomposition of $R(5, 11)^-$ into a left $R(5, 5)^-$ and a full right $R(5, 6)$ works for these 6 cases. Five of the remaining 11 cases are done in ad hoc fashion in FIGURE 9. A dark outlined 2×2 square appears in the tiling for the $(4, 1)$ case that is shown as the top left picture of FIGURE 9. Rotate that square 90° clockwise to produce a tiling for the $(4, 2)$ case; then rotate it another 90° to produce a tiling for the $(5, 2)$

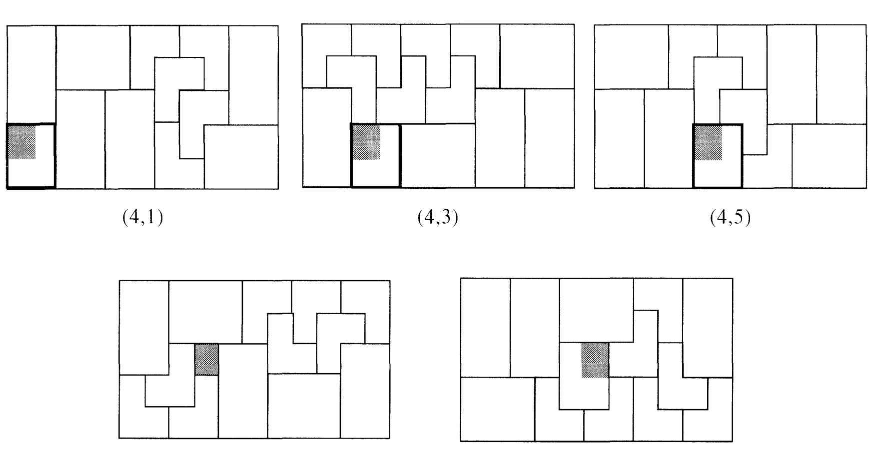


Figure 9 Deficient 5×11 tilings

case. Similar pairs of rotations produce tilings of the (4, 4) and (5, 4) cases from the displayed tiling of the (4, 3) case, as well as tilings of the (4, 6) and (5, 6) cases from the displayed tiling of the (4, 5) case. ■

Results and questions about 2-deficiency If two squares are removed from a rectangle, call the resultant shape a 2-*deficient rectangle*. The following proposition disallows the possibility of making the natural definition of deficiency of order k, $k \geq 2$ and then finding a direct extension of the Deficient Rectangle Theorem for higher deficiencies.

PROPOSITION. *No rectangle has the property that no matter which two* 1×1 *squares are removed, the remaining shape of area* $mn - 2$ *can be tiled.*

For if the squares $(1, 2)$ and $(2, 1)$ are removed, then the square $(1, 1)$ cannot be covered by a tromino. (This also shows the proposition still holds even if "tiling" is extended to mean "tiling by any collection of polyominoes which contains no monomino.")

Even though there will not be a direct analogue of the Deficient Rectangle Theorem, there is room for some interesting work to be done here. Here is a program for what to do about 2-deficiency. We extend the definition of good to 2-deficiency. A pair of squares is *good* if their removal from a $m \times n$ rectangle leaves a figure that can be tiled.

PROBLEM. *For the general case of* 2-*deficiency, find all bad pairs of squares for all* $m \times n$ *rectangles where* $mn \equiv 2 \pmod 3$. *Slightly less generally, exactly when can such a rectangle be covered by one domino and* $(mn - 2)/3$ *trominoes?*

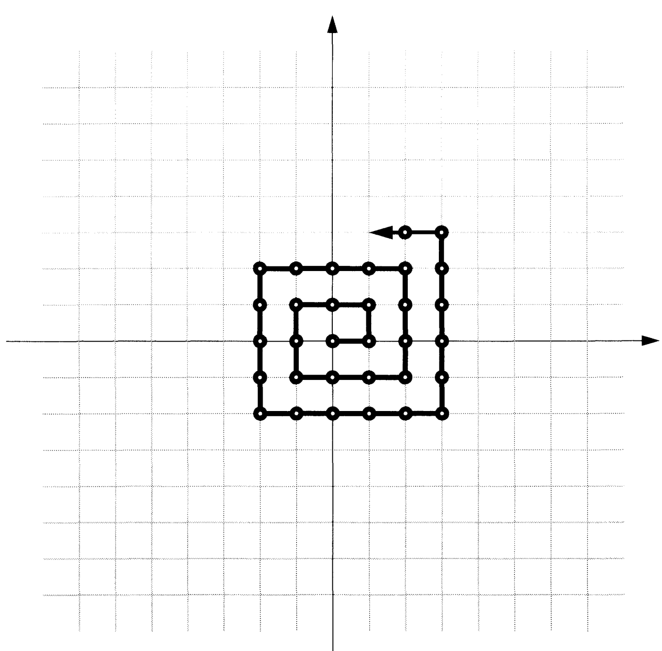
On the negative side, as we pointed out in the proof of the Proposition, the pair $\{(2, 1), (1, 2)\}$ is bad, that is, if square $(2, 1)$ and square $(1, 2)$ are removed, then no tromino can cover square $(1, 1)$. On the positive side, recall that in the application given after the proof of the Dog-eared Rectangle Theorem we pointed out that a tiling is always possible if the two removed squares are adjacent and in a corner of the rectangle. In other words, if $mn \equiv 2 (\text{mod } 3)$, then the pair $\{(1, n), (2, n)\}$ is good. Now consider the 5×7 case. As in the analysis of the 5×5 case for deficient rectangles done above, form a checkerboard-like pattern by marking each of the 12 squares that have both coordinates odd and assume that two of the 23 unmarked squares have been removed from $R(5, 7)$ to form $R^{=}$. Then a proposed tiling of $R^{=}$ must contain one tromino for each of the 12 marked squares, so that tiling must have area at least $12 \cdot 3 = 36$, which is absurd since the area of $R^{=}$ is 33. This reasoning disqualifies $\binom{23}{2} = 253$ pairs. Similar reasoning identifies a large number of bad pairs for $R(5, 13), \ldots, R(5, 7 + 6k), \ldots$.

Acknowledgments. Marshall Ash had profitable discussions with Ayse Sahin and Roger Jones. Martin Gardner provided encouragement and some references. Elizabeth Breuer motivated the application given after the proof of the Dog-eared Rectangle Theorem. Jerrold Griggs and a referee provided helpful historical information. We thank the referees for style-improving suggestions. This research was partially supported by NSF grant DMS 9707011 and a grant from the Faculty and Development Program of the College of Liberal Arts and Sciences, DePaul University.

REFERENCES

1. I-P. Chu and R. Johnsonbaugh, Tiling deficient boards with trominoes, this MAGAZINE **59** (1986), 34–40, MR 87c:05044.
2. ———, Tiling boards with trominoes, *J. Rec. Math.* **18** (1985-86), 188–193.
3. S. W. Golomb, Checker Boards and Polyominoes, *Amer. Math. Monthly*, **61** (1954), 675–682, MR 16,644a.
4. ———, *Polyominoes. Puzzles, patterns, problems, and packings*, 2nd ed., Princeton University Press, Princeton, 1994, MR 95k:00006.
5. B. Grünbaum and G. C. Shephard, *Tilings and Patterns*, W. H. Freeman, New York, 1987, MR 88k:52018.
6. R. Johnsonbaugh, *Discrete Mathematics*, 5th ed., Prentice Hall, Upper Saddle River, NJ, 2001.
7. *Amer. Math. Monthly* Problem 10641, Proposed by J. R. Griggs, **105** (1998), 175; solution by Michael Woltermann, **107** (2000), 179.
8. R. Penrose, *The Emperor's New Mind*, Oxford University Press, New York, 1989, MR 91c:00010.
9. P. Zeitz, *The Art and Craft of Problem Solving*, John Wiley, New York, 1999, MR 2000f:00001.

Proof Without Words: $\mathbb{Z} \times \mathbb{Z}$ Is a Countable Set



—DES MACHALE
UNIVERSITY COLLEGE
CORK, IRELAND

Extremal Curves of a Rotating Ellipse

CARL V. LUTZER
JAMES E. MARENGO
Rochester Institute of Technology
Rochester, NY 14623
cvlsma@rit.edu

Denote by $N(\phi)$ the highest point on an ellipse after it's been rotated through an angle ϕ about the origin. At first glance, it might appear that $N(\phi)$ parametrizes an ellipse that is parallel and tangent to the original (FIGURE 1). Is it actually an ellipse? We answer this question by finding the horizontal lines that intersect the rotated ellipse exactly once and using this information to eliminate the parameter, ϕ. The remainder of this exposition is devoted to answering other geometric questions arising in this context. We find the Cartesian equation that describes the paths traced out by the right-most and left-most points of the rotated ellipse ($E(\phi)$ and $W(\phi)$), and we study the line segments connecting these points. The relationship between the original ellipse and the curves parametrized by N, S, E, and W is investigated, and the golden mean makes a surprise appearance.

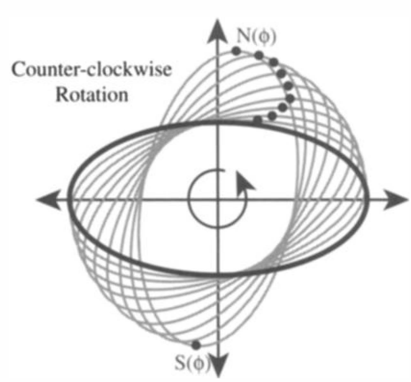


Figure 1 Tracking the high point

Description of the extremal curves Choose $a > c > 0$ and set $b^2 = a^2 - c^2$. Then the ellipse $x^2/a^2 + y^2/b^2 = 1$ is centered at the origin, its major axis is parallel to the x-axis, and its foci are at $(\pm c, 0)$. If the ellipse rotates through ϕ radians about the origin, its foci will be at $\pm(c\cos(\phi), c\sin(\phi))$, and all points (x, y) on the ellipse satisfy

$$\sqrt{(x - c\cos(\phi))^2 + (y - c\sin(\phi))^2} + \sqrt{(x + c\cos(\phi))^2 + (y + c\sin(\phi))^2} = 2a. \tag{1}$$

With a little algebra, we can rewrite (1) as

$$(a^2 - c^2\cos^2\phi)x^2 - (2c^2\cos\phi\sin\phi)xy + (a^2 - c^2\sin^2\phi)y^2 = a^2b^2. \tag{2}$$

Of course, when $\phi = 0$, this reduces to $x^2/a^2 + y^2/b^2 = 1$.

The key to finding the locus of the moving point $N(\phi)$ is the fact that a horizontal line through $N(\phi)$ (or though the lowest point, $S(\phi)$) intersects the ellipse exactly once. Hence, we examine the intersection of the rotated ellipse with lines of the form $y = h$ by substituting $y = h$ into (2). This yields a quadratic equation in x with discriminant $\Delta = 4a^2b^2(a^2 - c^2\cos^2\phi - h^2)$. This has a unique solution exactly when $\Delta = 0$, which means $h = \pm\sqrt{a^2 - c^2\cos^2\phi}$. After determining the corresponding values of x from (2), we see that

$$N(\phi) = \left(\frac{c^2\cos\phi\sin\phi}{\sqrt{a^2 - c^2\cos^2\phi}} \, , \, \sqrt{a^2 - c^2\cos^2\phi} \right), \tag{3}$$

and $S(\phi) = -N(\phi)$. To find the Cartesian equation of the curves parametrized by N and S, we eliminate the parameter from (3). Writing $N(\phi) = (x, y)$, we do a little algebra to find an equation for the desired locus,

$$x^2y^2 = (a^2 - y^2)(y^2 - b^2). \tag{4}$$

Similarly, one can find the right-most and left-most points, $E(\phi)$ and $W(\phi)$, by intersecting vertical lines with the rotated ellipse. Eliminating the parameter from the resulting coordinate expression for $E(\phi)$ yields

$$x^2y^2 = (a^2 - x^2)(x^2 - b^2). \tag{5}$$

The reader will note that equation (4) answers our earlier question: the curve parametrized by N is *not* an ellipse because its equation has degree 4. But then what kind of curve is it? The curve parametrized by N is symmetric with respect to the y-axis, and its y-intercepts are at $y = a$ and $y = b$. Consequently, the only horizontal line about which this curve could be symmetric is $y = (a + b)/2$. We investigate the possibility of symmetry by finding the curve's right-most point. If this extremal point is unique, it will have to lie on the line of symmetry. Begin by rewriting equation (4) as

$$x^2 = (a - b)^2 - \left(y - \frac{ab}{y} \right)^2. \tag{6}$$

From equation (6) it's clear that $|x|$ is largest when $y = \sqrt{ab}$. Since $0 < b < a$, the geometric mean is below the arithmetic mean, so the curve parametrized by N is not symmetric about *any* horizontal line. The same argument applies to the curve parametrized by S.

The curves parametrized by N and S will always be tangent to, and outside the original, unrotated ellipse. However, the relationship between the original ellipse and the E and W extremal curves is more interesting. In particular, when is the highest point of (5) inside, on, or outside the original, unrotated ellipse?

Intersections of the unrotated ellipse and the extremal curve parametrized by E, if there are any, can be found by combining (5) and the equation $x^2/a^2 + y^2/b^2 = 1$, where $x > 0$. We solve the latter for y^2 and substitute into (5) to find

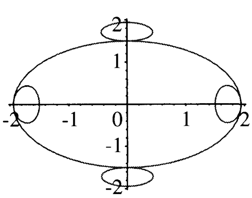
$$x^2\frac{b^2}{a^2}(a^2 - x^2) = (a^2 - x^2)(x^2 - b^2)$$

so that $x = a$ or $x = ab/\sqrt{a^2 - b^2}$. We know that $x \le a$ so, if these two curves intersect at some point other than $(a, 0)$, it must be that $ab/\sqrt{a^2 - b^2} < a$, whence $b < a/\sqrt{2}$. We conclude that the extremal curve parametrized by E intersects the unrotated ellipse at $(a, 0)$ and

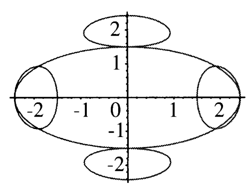
$$\left(\frac{ab}{\sqrt{a^2-b^2}}, \pm\frac{b\sqrt{a^2-2b^2}}{\sqrt{a^2-b^2}}\right). \tag{7}$$

Note that (7) reduces to $(a, 0)$ exactly when $b = a/\sqrt{2}$.

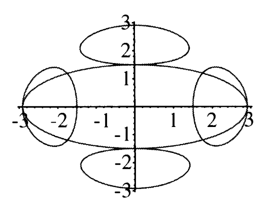
Equation (5) can be written in the form of (6), from which it follows that the highest and lowest points of the right extremal curve occur at $(\sqrt{ab}\ ,\ \pm(a-b))$. Comparing these coordinates to (7), we see that the highest and lowest points occur inside, on, or outside the original, unrotated ellipse according as a/b is less than, equal to, or greater than $(\sqrt{5}+1)/2$, the golden mean.



$\frac{a}{b} < \frac{1+\sqrt{5}}{2}$



$\frac{a}{b} = \frac{1+\sqrt{5}}{2}$



$\frac{a}{b} > \frac{1+\sqrt{5}}{2}$

Extremal secants After a nontrivial rotation, the line segments that connect the highest to the lowest, and the right-most to the left-most point of the ellipse do not remain orthogonal. How close together do they come?

We will designate the line segment connecting the highest to the lowest as l_{ns} and the line segment connecting the right-most to the left-most point as l_{ew}. Because l_{ns} and l_{ew} pass through the origin, equation (3) tells us that

$$l_{ns} = \left\{(x, y) : (c^2\cos\phi\sin\phi)y = (a^2 - c^2\cos^2\phi)x\right\}. \tag{8}$$

Equations describing l_{ew} are similar. Let us designate the angle between l_{ew} and l_{ns} by θ. In order to express θ in terms of the rotation angle, ϕ, we use the formula for the tangent of the difference of two angles. After some algebra, with the slope of l_{ns} as $\tan\alpha$ and the slope of l_{ew} as $\tan\beta$, this identity gives us,

$$\theta = \tan^{-1}\left(\frac{2a^2b^2}{(a^4-b^4)\sin(2\phi)}\right).$$

This angle is minimized when $\phi = \frac{\pi}{4}$. Note that, for fixed $\phi \in (0, \frac{\pi}{2})$, θ increases monotonically from 0 to $\frac{\pi}{2}$ as b increases from 0 to a.

Furthermore, if θ_1 and θ_2 are the angles formed by the major axis of the ellipse with the extremal secants l_{ew} and l_{ns}, respectively, one can show that

$$\theta_1 = \tan^{-1}\left(\frac{b^2\tan\phi}{a^2}\right) \quad \text{and} \quad \theta_2 = \tan^{-1}\left(\frac{b^2}{a^2\tan\phi}\right).$$

Of course, $\theta_1 + \theta_2 = \theta$. Also, if $a = b$, we see $\theta_1 = \phi$ and $\theta_2 = \frac{\pi}{2} - \phi$, as expected.

It is interesting to note that, before rotation, l_{ns} (resp. l_{ew}) bisects any horizontal (resp. vertical) secant of the ellipse. This property is retained after rotation, as the following calculation shows: Suppose $-b < h < b$. By substituting $y = h$ into (2) and solving the resulting quadratic equation for x, we find that the points on the ellipse

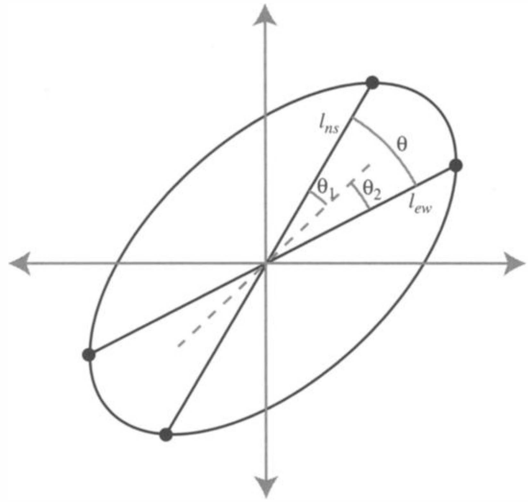


Figure 2 Extremal secants

with ordinate y have abscissa

$$x = \frac{(c^2 \cos\phi \sin\phi)h \pm ab\sqrt{a^2 - c^2\cos^2\phi - h^2}}{a^2 - c^2\cos^2\phi}. \tag{9}$$

If we also substitute $y = h$ into the defining equation of l_{ns} (see 8), we find that

$$x = \frac{hc^2 \cos\phi \sin\phi}{a^2 - c^2\cos^2\phi},$$

and this value is exactly half way between the values calculated in (9). A similar calculation shows that l_{ew} bisects any vertical secant of the ellipse. In fact, since ϕ is arbitrary, if parallel lines L_1 and L_2 are tangent to the ellipse at P_1 and P_2, respectively, the secant of the ellipse connecting P_1 and P_2 bisects any secant that is parallel to L_1.

A connection to probability Suppose X and Y represent the heights of the husband and wife in a married couple chosen randomly from a population of such couples. It may of interest to consider the distribution of heights of wives who are married to men of a specified height. This distribution is called a *conditional distribution of Y given the value of X,* and its mean is referred to as a *conditional mean.*

We will denote the probability that $X \in [x, x + \Delta x]$ *and* $Y \in [y, y + \Delta y]$ by $P(x < X < x + \Delta x, y < Y < y + \Delta y)$. Suppose that $P(x < X < x + \Delta x, y < Y < y + \Delta y) \approx f(x, y)\Delta x \Delta y$ and that equality is achieved in the limit as Δx and Δy tend to zero. Then we say that f is the *joint probability density function* of the random vector (X, Y).

Random vectors such as those from the above example are often modeled by the *bivariate normal distribution.* The joint probability density function of this distribution is given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{g(x,y)},$$

where

$$g(x, y) = \frac{-\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}{2(1-\rho^2)}$$

and μ_x, μ_y, σ_x^2, σ_y^2, and ρ are the mean of X, mean of Y, variance of X, variance of Y, and the correlation coefficient of X and Y respectively. (The correlation coefficient is a measure of the linear association between X and Y.) The graph of f is a bell-shaped surface (FIGURE 3) and, for any $k \in (0, \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}]$, the level curve $f(x, y) = k$ is given by the ellipse

$$\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 = \ell, \tag{10}$$

where $\ell = -2(1-\rho^2)\ln(2\pi k\sigma_x\sigma_y\sqrt{1-\rho^2})$. FIGURE 3 depicts the graph of f and its intersection with a pair of horizontal planes (which result in the ellipses described by equation (10)) when $\mu_x = 64$, $\mu_y = 69$, $\sigma_x = \sigma_y = 3$, $\rho = 0.6$. It is well known [**1**] that the conditional means of this joint probability distribution are linear (the so-called population regression lines) and are given by

$$\mathcal{E}_X(Y) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(X - \sigma_x) \quad \text{and} \quad \mathcal{E}_Y(X) = \mu_x + \frac{\rho\sigma_x}{\sigma_y}(Y - \mu_y).$$

It is easily checked, using arguments similar to those given previously, that $\mathcal{E}_X(Y)$ intersects the level curve described by (10) at its right-most and left-most points, and $\mathcal{E}_Y(X)$ intersects this level curve at its highest and lowest points. That is, the graphs of $\mathcal{E}_X$ and $\mathcal{E}_Y$ are exactly the extremal secants of the rotated ellipse!

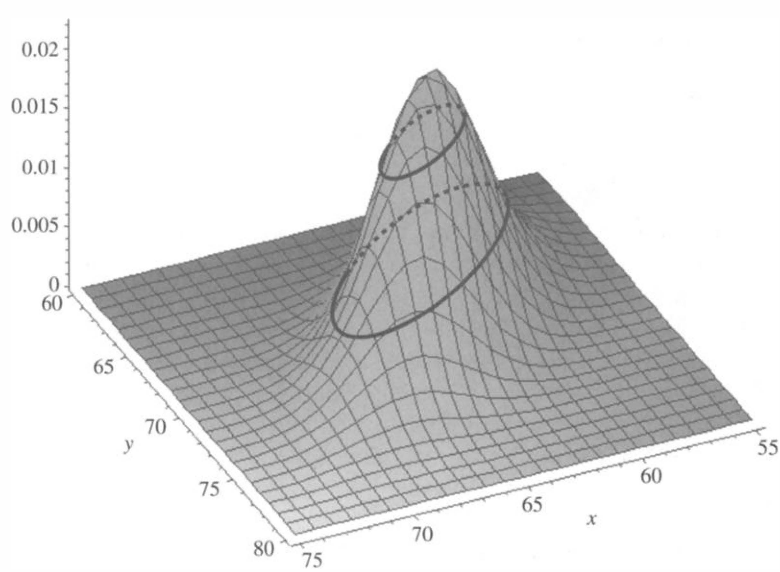


Figure 3 The graph of $z = f(x, y)$

Acknowledgments. The authors wish to thank Dr. David Farnsworth for his helpful comments.

REFERENCE

1. B. W. Lindgren, *Statistical Theory*, 3rd. ed., Macmillan, NY, 1976.

A Question of Limits

PAUL SCHUETTE
Meredith College
3800 Hillsborough Street
Raleigh, North Carolina 27607-5298
schuette@meredith.edu

While reviewing a draft of an assignment I was about to give to a multivariable calculus class, it occurred to me that all of the limit problems involved rational functions. In a moment of what I would like to call inspiration, I decided to add a twist to a familiar problem. A standard topic in the first semester of calculus is a demonstration that

$$\lim_{x\to 0}\frac{\sin x}{x} = 1.$$

A natural generalization of this limit is

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}}\frac{\sin x+\sin y}{x+y}. \tag{1}$$

The assignment I gave to my calculus class was to determine if (1) exists and to evaluate the limit if it does.

The purpose of this note is to discuss why is (1) an interesting limit, and to suggest that a broader range of problems can be investigated by students at various levels. Limits are one of the staples of single-variable calculus courses, yet the treatment of limits in multivariable calculus tends to be rather minimal. When considering singularities, many standard texts such as Stewart [**6**] deal almost exclusively with rational functions. This is also the case with many advanced calculus texts [**2, 4, 8**]. Undoubtedly, the increased complexity of limits in multiple dimensions partially accounts for the sparse treatment. However, there are many interesting mathematical questions regarding multivariable limits suitable for exploration by undergraduates.

How does one evaluate the limit in (1), provided, of course, that it even exists? Some of my students suggested that a graph such as FIGURE 1(a), produced by *Maple*, suffices. On the other hand, another software package, *Matlab*, produced the image in FIGURE 1(b) (the contours of the graph are plotted in the xy plane). This definitely gives a different view of the limit.

Given the discrepancies in the images, it isn't clear that either figure suffices to show, even at an intuitive level, that (1) exists.

There are several rigorous ways to evaluate the limit. One of the most elegant is to make the change of variables $x = u + v$ and $y = u - v$. Then

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}}\frac{\sin x+\sin y}{x+y} = \lim_{\substack{(u,v)\to(0,0)\\ u\neq 0}}\frac{\sin(u+v)+\sin(u-v)}{2u}$$

$$= \lim_{\substack{(u,v)\to(0,0)\\ u\neq 0}}\frac{2\sin(u)\cos(v)}{2u} = 1. \tag{2}$$

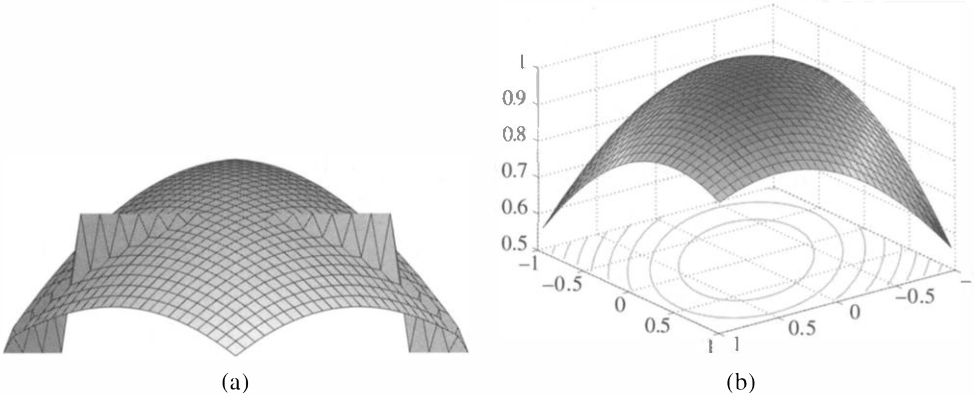


Figure 1 Two renderings of the same function graph

Another possibility is to let $z = -y$, so

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{\sin x + \sin y}{x+y} = \lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{\sin x - \sin(-y)}{x-(-y)}$$
$$= \lim_{\substack{(x,z)\to(0,0)\\ x\neq z}} \frac{\sin x - \sin z}{x-z}$$
$$= \sin'(0) = \cos(0) = 1. \qquad (3)$$

Equating the limit with the derivative is natural, but does require some justification (see Theorem 3 and Example 3). It may be noted that the method of (2) employs the addition formula for the sine, while the method of (3) uses only the fact that the sine function is odd and a definition of the derivative.

Generalizing the result Based on the results of (2) and (3), one might suppose that (1) could be generalized in a fairly obvious way.

A Question: *If* $\lim_{x\to 0} \frac{f(x)}{x} = a$, *does*

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{f(x)+f(y)}{x+y} = a? \qquad (4)$$

We can observe that both of the iterated limits are equal to a:

$$\lim_{x\to 0}\lim_{y\to 0} \frac{f(x)+f(y)}{x+y} = \lim_{x\to 0} \frac{f(x)+0}{x+0} = \lim_{y\to 0} \frac{f(y)+0}{y+0} = \lim_{y\to 0}\lim_{x\to 0} \frac{f(x)+f(y)}{x+y}.$$

This might suggest that it would be relatively simple to prove that the limit is a. However, if $f(x) = x^2$, the limit in (4) does not exist, as the following counterexample shows.

COUNTEREXAMPLE. $\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{x^2+y^2}{x+y}$ *does not exist.*

Proof. Consider the case when $x = 0$, so that one approaches the origin along the y-axis. Then $\lim_{y\to 0} \frac{y^2}{y} = 0$.

Now let $x = t^2 + t$ and $y = t^2 - t$. Then, provided $t \neq 0$,

$$\frac{x^2 + y^2}{x + y} = \frac{t^2(t+1)^2 + t^2(t-1)^2}{2t^2} = \frac{1}{2}\left[(t+1)^2 + (t-1)^2\right]$$

Since $\lim_{t\to 0} \frac{1}{2}\left[(t+1)^2 + (t-1)^2\right] = 1 \neq 0$, this establishes the counterexample. ∎

Observe that a graph of $z = \frac{x^2+y^2}{x+y}$ near the origin, such as the one produced with *Matlab* in FIGURE 2, doesn't demonstrate conclusively whether the limit exists or not. This example demonstrates that one cannot always trust the output of computer packages.

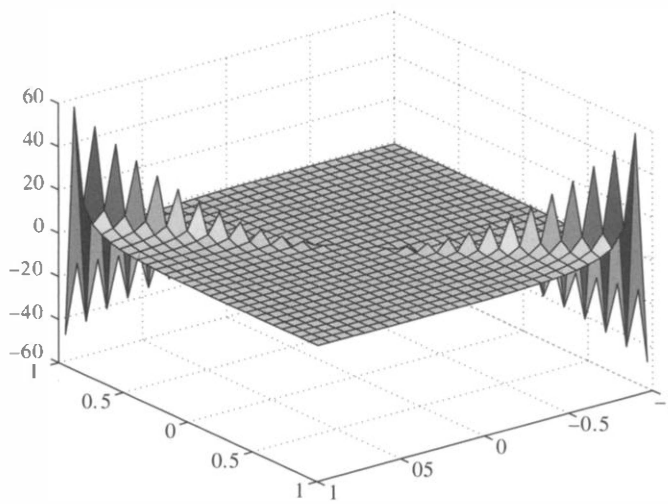


Figure 2 A misleading image

So when does (4) hold? Observe that $\sin(x)$ is odd, while the function $f(x) = x^2$ is even. This suggests the following somewhat surprising result.

THEOREM 1. *If* $\lim_{x\to 0} \frac{f(x)}{x} = a_1$, *and* f *is real analytic at 0, then*

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{f(x) + f(y)}{x + y} = a_1 \qquad (5)$$

if and only if f *is an odd function.*

The proof relies on the fact that $f(x)$ may be represented as a power series, that $x^{2k+1} + y^{2k+1}$ is divisible by $x + y$, and that we may generalize our counterexample.

Proof. Observe that $\lim_{x\to 0} f(x)/x = a_1$ implies $f(0) = 0$. Since f is real analytic, for some $r > 0$, $f(x) = \sum_{n=1}^{\infty} a_n x^n$ whenever $|x| < r$.

Suppose f is an odd function. Then $a_{2k} = 0$, for all $k \in \mathbb{N}$. Thus,

$$f(x) = \sum_{k=0}^{\infty} a_{2k+1} x^{2k+1},$$

for $|x| < r$, and hence, for $|x| < r$, $|y| < r$,

$$\frac{f(x)+f(y)}{x+y} = \sum_{k=0}^{\infty} a_{2k+1} \frac{x^{2k+1}+y^{2k+1}}{x+y}.$$

Let $p_{2k}(x, y) = x^{2k} - x^{2k-1}y + \ldots - xy^{2k-1} + y^{2k}$. Observe that if $x + y \neq 0$, then $p_{2k}(x, y) = (x^{2k+1} + y^{2k+1})/(x + y)$. Thus,

$$\frac{f(x)+f(y)}{x+y} = a_1 + \sum_{k=1}^{\infty} a_{2k+1}\, p_{2k}(x, y). \tag{6}$$

If $\rho < r$, with $|x| < \rho/2$ and $|y| < \rho/2$, then

$$\left|p_{2k}(x, y)\right| \leq (2k+1)\left(\frac{\rho}{2}\right)^{2k} = \frac{(2k+1)\rho^{2k}}{4^k}.$$

Because $0 < (2k+1)/4^k < 1$ and $\rho < r$, $\sum_{k=1}^{\infty} |a_{2k+1}|(2k+1)\rho^{2k}/4^k < \infty$. Therefore, by a version of the Weierstrass M-test [**7**, p. 141], $\sum_{k=1}^{\infty} a_{2k+1}\, p_{2k}(x, y)$ is absolutely convergent. Since $p_{2k}(x, y)$ is continuous in $\mathbb{R}^2$, with $p_{2k}(0, 0) = 0$, it follows that

$$\lim_{(x,y)\to(0,0)} \sum_{k=1}^{\infty} a_{2k+1} p_{2k}(x, y) = 0. \tag{7}$$

Thus, (5) follows from (6) and (7).

Now assume that (5) is true. Observe that if $|x| < r$ and $|y| < r$,

$$\frac{f(x)+f(y)}{x+y} = a_1 + \sum_{k=1}^{\infty} a_{2k+1}\, p_{2k}(x, y) + \sum_{k=1}^{\infty} a_{2k} \frac{x^{2k}+y^{2k}}{x+y}.$$

From (5) and (7), it follows that

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \sum_{k=1}^{\infty} a_{2k} \frac{x^{2k}+y^{2k}}{x+y} = 0. \tag{8}$$

The proof of the theorem will be complete with the following claim.

CLAIM. *If the limit in (8) is* 0, *then* $a_{2k} = 0$ *for all* $k \in \mathbb{N}$.

Proof of the claim: Let k_0 be the least value of k for which $a_{2k} \neq 0$. Modifying the approach of our earlier counterexample, for $t > 0$, let $x(t) = t^\alpha + t$ and $y(t) = t^\alpha - t$, where $\alpha > 0$ will be determined by k_0. Then, $(x^{2k} + y^{2k})/(x + y) = g_{2k}(t)$, where

$$g_{2k}(t) = \frac{t^{2k-\alpha}}{2}\left[\left(t^{\alpha-1}+1\right)^{2k} + \left(t^{\alpha-1}-1\right)^{2k}\right].$$

From (8) it follows that $\lim_{t\to 0^+} \sum_{k=1}^{\infty} a_{2k}\, g_{2k}(t) = 0$. However, if $\alpha = 2k_0$,

$$\sum_{k=1}^{\infty} a_{2k}\, g_{2k}(t) = \frac{a_{2k_0}}{2}\left[\left(t^{2k_0-1}+1\right)^{2k_0} + \left(t^{2k_0-1}-1\right)^{2k_0}\right] + \sum_{k=k_0+1}^{\infty} a_{2k}\, g_{2k}(t). \tag{9}$$

For $t > 0$, $0 < g_{2k}(t) \leq ((t + t^\alpha)^{2k})/t^\alpha$. Since $\lim_{t\to 0^+}((t + t^\alpha)^{2k})/t^\alpha = 0$ uniformly in k for $k > k_0$, the absolute convergence of $\sum_{k=k_0+1}^{\infty} a_{2k}\, x^{2k}$ in a neighborhood of 0 implies

$$\lim_{t\to 0^+} \sum_{k=k_0+1}^{\infty} a_{2k} g_{2k}(t) = 0.$$

From (9), we may conclude that

$$\lim_{t\to 0^+} \sum_{k=1}^{\infty} a_{2k}\, g_{2k}(t) = \lim_{t\to 0^+} \frac{a_{2k_0}}{2} \left[\left(t^{2k_0-1}+1\right)^{2k_0} + \left(t^{2k_0-1}-1\right)^{2k_0} \right] = a_{2k_0}.$$

This, however, contradicts (8). Thus, $a_{2k} = 0$ for all $k \in \mathbb{N}$. ∎

EXAMPLE 1. If $f(x) = \sin(x)$, then Theorem 1 immediately yields

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{\sin x + \sin y}{x+y} = 1.$$

EXAMPLE 2. If $f(x) = e^x - 1$, then $f(0) = 0$, $\lim_{x\to 0} f(x)/x = 1$, but

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{e^x + e^y - 2}{x+y}$$

doesn't exist, because f is not odd.

The strong derivative Theorem 1 states that if f is real analytic and the other hypotheses hold, then

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{f(x)+f(y)}{x+y}$$

exists only if f is odd. If f is an odd function, can the condition that f is real analytic be relaxed? We shall address this question in Theorem 3 below, which employs an alternative definition of the derivative.

Suppose that f is an odd function. As in (3), let $z = -y$. Then

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{f(x)+f(y)}{x+y} = \lim_{\substack{(x,z)\to(0,0)\\ x\neq z}} \frac{f(x)-f(z)}{x-z}.$$

This leads to the definition of the *strong derivative* $f^*(x_0)$ by

$$f^*(x_0) = \lim_{\substack{(x,z)\to(x_0,x_0)\\ x\neq z}} \frac{f(x)-f(z)}{x-z},$$

when the limit exists. Bruckner and Leonard [**1**] attribute the definition of the strong derivative to Peano [**5**], who is well known for his axioms for the natural numbers. Esser and Shisha [**3**] show that if $f^*(x)$ exists, then $f^*(x) = f'(x)$; that $f^*(x)$ is continuous on its domain of definition; and provide necessary and sufficient conditions for the existence of the strong derivative. They also give an easily checked sufficient condition for the existence of the strong derivative [**3**]:

THEOREM 2. *If f' is continuous at a point x_0, then f is strongly differentiable at x_0.*

The definition of the strong derivative and Theorem 2 immediately yield Theorem 3, which extends the results of Theorem 1 in one direction.

THEOREM 3. *If f is an odd function, $f(0) = 0$, $f'(0) = a$ and $f'(x)$ is continuous at 0, then*

$$\lim_{\substack{(x,y)\to(0,0)\\ x+y\neq 0}} \frac{f(x)+f(y)}{x+y} = f^*(0) = f'(0) = a.$$

EXAMPLE 3. If $f(x) = \sin(x)$ then (2) follows immediately from Theorem 3. Theorem 2 may also be applied at points of the form $(x_0, -x_0)$, yielding

$$\lim_{\substack{(x,y)\to(x_0,-x_0)\\ x+y\neq 0}} \frac{\sin(x)+\sin(y)}{x+y} = \cos(x_0).$$

Hence, the function

$$g(x,y) = \begin{cases} \dfrac{\sin(x)+\sin(y)}{x+y} & \text{if } x \neq -y \\ \cos(x), & \text{if } x = -y \end{cases}$$

is a continuous extension to all of $\mathbb{R}^2$ of $z = \frac{\sin(x)+\sin(y)}{x+y}$. The function $g(x,y)$ can be shown to be differentiable, since the partial derivatives exist and are continuous. Thus, FIGURE 1(b) is a more accurate depiction of the behavior of the function than FIGURE 1(a).

EXAMPLE 4. If $f(x) = \cos(x)$, then Theorem 2 yields

$$\lim_{\substack{(x,y)\to(0,0)\\ x\neq y}} \frac{\cos(x)-\cos(y)}{x-y} = -\sin(0) = 0.$$

This isn't exactly obvious from a graph such as FIGURE 3.

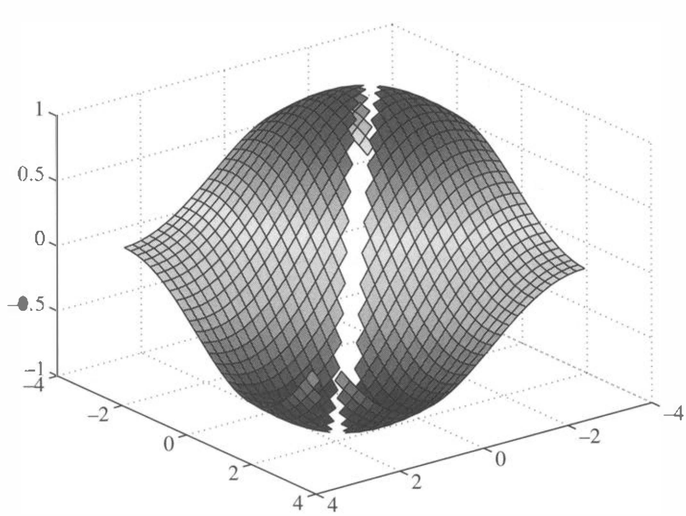


Figure 3 An uninformative image

Higher dimensions Since we can evaluate (1) as a limit in two dimensions, it is natural to inquire whether similar results hold in higher dimensions. For example, does

$$\lim_{\substack{(x,y,z)\to(0,0,0)\\ x+y+z\neq 0}} \frac{\sin(x)+\sin(y)+\sin(z)}{x+y+z} = 1? \tag{10}$$

Interestingly, the limit on the left-hand side of (10) does not exist, as the following theorem shows.

THEOREM 4. *If* $\lim_{x\to 0}\frac{f(x)}{x} = a_1$, *and* f *is real analytic at 0, then*

$$\lim_{\substack{(x,y,z)\to(0,0,0)\\ x+y+z\neq 0}} \frac{f(x)+f(y)+f(z)}{x+y+z} = a_1 \tag{11}$$

if and only if $f(x) = a_1 x$.

Much of the proof of Theorem 4 is analogous to Theorem 1, so only a brief sketch of the proof will be provided.

First, if $f(x) = a_1 x$, then (11) follows trivially. Now assume that (11) holds, and that $f(x) = \sum_{n=1}^{\infty} a_n x^n$ in a neighborhood of 0. If $z = 0$, then (11) reduces to (5), which allows us to conclude that $a_{2k} = 0$ for all $k \in \mathbb{N}$. Let $x = y = t^3 - t$, and $z = t^3 + 2t$, for $t > 0$. Then

$$\frac{f(x(t))+f(y(t))+f(z(t))}{x(t)+y(t)+z(t)} = a_1 + \frac{a_3}{3}\left(2(t^2-1)^3 + (t^2+2)^3\right) + \sum_{k=2}^{\infty} a_{2k+1}\, h_{2k+1}(t),$$

where

$$h_{2k+1}(t) = \frac{2(t^3-t)^{2k+1} + (t^3+2t)^{2k+1}}{3t^3}.$$

As in the proof of Theorem 1, it may be shown that $\lim_{t\to 0^+}\sum_{k=2}^{\infty} a_{2k+1}\, h_{2k+1}(t) = 0$, leaving

$$\lim_{t\to 0^+} \frac{f(x(t))+f(y(t))+f(z(t))}{x(t)+y(t)+z(t)} = a_1 + 2a_3.$$

However, if (11) holds, then $a_3 = 0$. A similar argument, changing the highest power in the parameterization of x, y and z, shows $a_5 = 0$, and so on. ■

From Theorem 4, we immediately obtain

THEOREM 5. *If* $\lim_{x\to 0} f(x)/x = a_1$ *and* f *is real analytic at 0, then, for integer* $n \geq 3$,

$$\lim_{\substack{(x_1,x_2,\ldots,x_n)\to(0,0,\ldots,0)\\ x_1+x_2+,\ldots,x_n\neq 0}} \frac{f(x_1)+f(x_2)+\ldots+f(x_n)}{x_1+x_2+,\ldots,x_n} = a_1$$

if and only if $f(x) = a_1 x$.

Conclusions and suggestions Limits in $\mathbb{R}^2$ and $\mathbb{R}^n$ can be a source of interesting and sometimes counterintuitive problems. Many of the results of this note can be readily used at the undergraduate level. Students in a multivariable calculus course may certainly be asked to evaluate (1), possibly with aspects of (2) and (3) given as hints. Using the counterexample as a model, students may be guided to conjecture Theorem 1, although the proof of Theorem 1 would be more appropriate for a course in advanced calculus or elementary analysis. The strong derivative can be used to revisit the definition of the derivative, and build on earlier concepts of the meaning of the derivative. To show that Theorem 1 cannot be extended to higher dimensions, the parameterization given in the outline of the proof of Theorem 4 may be used to show that the limit of the left-hand side of (10) doesn't exist.

Acknowledgments. The author would like to thank Peter Jarvis, Warren Johnson, and the referees for their many useful and helpful suggestions.

REFERENCES

1. A. M. Bruckner and J. L. Leonard, Derivatives, *Amer. Math. Monthly* **73** (1966), 24–56.
2. R. C. Buck with E. F. Buck, *Advanced Calculus*, 3rd ed., McGraw-Hill, New York, NY, 1978.
3. M. Esser and O. Shisha, A Modified Differentiation, *Amer. Math. Monthly* **71** (1964), 904–906.
4. W. Fulks, *Advanced Calculus*, 3rd ed., John Wiley, New York, NY, 1978.
5. G. Peano, Sur la définition de la dérivée, *Mathesis* **2** (1894), 12–14, (*Opere scelte*, vol. 1, Edizioni Cremonese, Roma, 1957, pp. 210–212).
6. J. Stewart, *Calculus*, 4th ed., Brooks/Cole, Pacific Grove, CA, 1999.
7. K. R. Stromberg, An Introduction to Classical Real Analysis, Wadsworth, Belmont, CA, 1981.
8. A. E. Taylor and W. R. Mann, *Advanced Calculus*, 3rd ed., John Wiley, New York, NY, 1983.

70 Years Ago in the MAGAZINE (called, at that time, *The Mathematics Newsletter*)

From "Improving the Teaching of College Mathematics," by May M. Beenken, Vol. **8**, No. 5, (Feb., 1934), 97–103:

> In order to keep mentally alert, the college teacher of mathematics should himself [*sic*] be working and learning constantly. We may well harken to the words of J. W. Young in his retiring presidential address to the Mathematical Association of America. He said, "The sin of the mathematician is not that he [*sic*] doesn't do research, the sin is idleness, when there is work to be done. If there be sinners in my audience, I would urge them to sin no more. If your interest is in research, do that; if you are of a philosophical temperament, cultivate the gardens of criticism, evaluation, and interpretation; if your interest is historical, do your plowing in the field of history; if you have the insight to see simplicity in apparent complexity, cultivate the field of advanced mathematics from the elementary point of view; if you have the gift of popular exposition, develop your abilities in that direction; if you have executive and organizing ability, place that ability at the disposal of your organization. Whatever your abilities there is work for you to do,—for the greater glory of mathematics." And may I add, "Whatever you do, do it for the greater glory of teaching, which is the chief purpose for which you are employed."

The editor hopes that those evaluating the scholarly achievement of faculty today will reward all the various types of endeavor advocated by Young.

PROBLEMS

ELGIN H. JOHNSTON, *Editor*
Iowa State University

Assistant Editors: RĂZVAN GELCA, Texas Tech University; ROBERT GREGORAC, Iowa State University; GERALD HEUER, Concordia College; VANIA MASCIONI, Ball State University; PAUL ZEITZ, The University of San Francisco

Proposals

To be considered for publication, solutions should be received by July 1, 2004.

1686. *Proposed by Shahin Amrahov, Ari College, Turkey.*

Find all positive integer solutions (x, y) to the equation

$$2y^2 = x^4 + 8x^3 + 8x^2 - 32x + 15.$$

1687. *Proposed by Sung Soo Kim, Hanyang University, Ansan Kyunggi, Korea.*

A two-player game starts with two sticks, one of length n and one of length $n + 1$, where n is a positive integer. Players alternate turns. A turn consists of breaking a stick into two sticks of positive integer lengths, or removing k sticks of length k for some positive integer k. The player who makes the last move wins. Which player can force a win?

1688. *Proposed by Mihai Manea, Princeton University, Princeton, NJ.*

Let p be an odd prime, and let $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{p-1}x^{p-1}$ be a polynomial of degree $p - 1$ with integral coefficients. Suppose that $p \nmid (P(b) - P(a))$ whenever a and b are integers such that $p \nmid (b - a)$. Prove that $p \mid a_{p-1}$.

1689. *Proposed by Ali Nabi Duman, student, Bilkent University, Ankara, Turkey.*

Triangle ABC is a right triangle with right angle at A. Circle $\mathcal{C}$ is tangent to $\overline{AB}$ and $\overline{BC}$ at K and N, respectively, and intersects $\overline{AC}$ in points $M(\neq A)$ and P, with $AM < AP$. The line perpendicular to $\overline{BC}$ at N intersects the median from A, the circle $\mathcal{C}$, and $\overline{AB}$ in points L, F, and E, respectively. Prove that if $FL/EF = LN/EN$, then

a. K, L, and M are collinear.
b. $\cos(2\angle ABC) = EA/EK$.

We invite readers to submit problems believed to be new and appealing to students and teachers of advanced undergraduate mathematics. Proposals must, in general, be accompanied by solutions and by any bibliographical information that will assist the editors and referees. A problem submitted as a Quickie should have an unexpected, succinct solution.

Solutions should be written in a style appropriate for this MAGAZINE. Each solution should begin on a separate sheet.

Solutions and new proposals should be mailed to Elgin Johnston, Problems Editor, Department of Mathematics, Iowa State University, Ames IA 50011, or mailed electronically (ideally as a LaTeX file) to ehjohnst@iastate.edu. All communications should include the readers name, full address, and an e-mail address and/or FAX number.

1690. *Proposed by Costas Efthimiou, Department of Physics, and Peter Hilton, Department of Mathematics, University of Central Florida, Orlando, FL.*

Prove that there exist functions $f : \mathbb{R} \to \mathbb{R}$ that satisfy

$$f\big(x - f(y)\big) = f(x) + y$$

for all $x,\ y \in \mathbb{R}$, and show how such functions can be constructed.

Quickies

Answers to the Quickies are on page 75.

Q937. *Proposed by Bill Chen, Philadelphia, PA, Clark Kimberling, Evansville, IN, and Paul R. Pudaite, Glen Ellyn IL.*

Let n be a positive integer. Prove that

$$\sum_{k=1}^{n} \left\lfloor \frac{n}{k} + \frac{1}{2} \right\rfloor - \sum_{k=1}^{n} \left\lfloor \frac{n}{k + \frac{1}{2}} \right\rfloor = n.$$

Q938. *Proposed by William P. Wardlaw, U. S. Naval Academy, Annapolis, MD.*

Let R be a ring, let G be a finite subset of R that forms a multiplicative group under the multiplication of R, and let s be the sum of the elements of G. Prove that if G has more than one element, then s is either zero or a zero divisor in R. Give examples in which s is a nonzero divisor of zero.

Solutions

A Square Bound **February 2003**

1662. *Proposed by Erwin Just (Emeritus) and Norman Schaumberger (Emeritus), Bronx Community College of the City of New York, Bronx, NY.*

Let x_k, $1 \le k \le n$, be positive real numbers with $\sum_{k=1}^{n} x_k^{2k-1} \le n$. Prove that $\sum_{k=1}^{n} (2k-1)x_k \le n^2$.

I. *Solution by Michael G. Neubauer, California State University, Northridge, CA.*

Bernoulli's Inequality states that if $r \ge 1$ and $x \ge 0$, then $x^r - 1 \ge r(x-1)$. Replace x by x_k and r by $2k-1$, then do some rearranging to obtain

$$(2k-1)x_k \le x_k^{2k-1} - 1 + (2k-1).$$

It follows that

$$\sum_{k=1}^{n}(2k-1)x_k \le \sum_{k=1}^{n} x_k^{2k-1} - n + \sum_{k=1}^{n}(2k-1) \le n - n + n^2 = n^2.$$

II. *Solution by Heinz-Jürgen Seiffert, Berlin, Germany.*

We prove the following generalization:

Let I be a real interval containing 1 and let $f_k : I \longrightarrow \mathbb{R}$, $1 \le k \le n$, be differentiable and convex on I. If c is a real number, and $x_k \in I$, $1 \le k \le n$, with

$\sum_{k=1}^{n} f_k(x_k) \le c$, then

$$\sum_{k=1}^{n} f_k'(1)x_k \le c + \sum_{k=1}^{n} \left(f_k'(1) - f_k(1)\right).$$

The result in the problem statement follows by taking $f_k(x) = x^{2k-1}$, $x_k \in I = (0, \infty)$, and $c = n$.

To establish the generalization, first observe that for $1 \le k \le n$, the function g_k defined by $g_k(x) = f_k'(1)x - f_k(x)$ satisfies

$$\begin{aligned} g_k'(x) &\ge 0 \qquad x \in I \text{ and } x < 1 \\ g_k'(x) &\le 0 \qquad x \in I \text{ and } x \ge 1, \end{aligned}$$

so $g_k(x) \le g_k(1)$ for all $x \in I$. Hence

$$\begin{aligned} \sum_{k=1}^{n} f_k'(1)x_k &\le c + \sum_{k=1}^{n} \left(f_k'(1)x_k - f_k(x_k)\right) \\ &= c + \sum_{k=1}^{n} g_k(x_k) \le c + \sum_{k=1}^{n} g_k(1) = c + \sum_{k=1}^{n} \left(f_k'(1) - f_k(1)\right). \end{aligned}$$

Also solved by Reza Akhlaghi, Tsehaye Andebrhan, Michael Andreoli, Carl Axness (Spain), Michel Bataille (France), Jean Bogaert (Belgium), Cal Poly Pomona Problem Solving Group, Minh Can, Mario Catalani (Italy), Con Amore Problem Group (Denmark), Knut Dale (Norway), Daniele Donini (Italy), Robert L. Doucette, Peter Drianov (Canada), Aaron Dutle, FGCU Problem Group, Ovidiu Furdui, G.R.A.20 Problems Group (Italy), Julien Grivaux (France), Enkel Hysnelaj (Australia), The Ithaca College Solvers, Steve Kaczkowski, Achim Kehrein (Germany), Murray S. Klamkin (Canada), Elias Lampakis (Greece), Kee-Wai Lau (China), Northwestern University Math Problem Solving Group, Albert D. Polimeni, Rob Pratt, Phillip P. Ray, Rolf Richberg (Germany), Joel Schlosberg, Harry Sedinger, Achilleas Sinefakopoulos, Nicholas C. Singer, John W. Spellmann, Nora Thornber, Dave Trautman, Chu Wenchang and Magli Pierluigi (Italy), Michael Vowe (Switzerland), John T. Zerger, Li Zhou, and the proposers.

Much Ado About Nothing **February 2003**

1663. *Proposed by Michel Bataille, Rouen, France.*

Let m and n be integers such that $1 \le m < n+1$. Evaluate

$$\sum_{k=1}^{n+1} \left((k+1)\sin^{k-1}\left(\frac{2\pi m}{n+1}\right) \prod_{j=1}^{k} \left(\cot\left(\frac{\pi m}{n+1}\right) - \cot\left(\frac{\pi j}{k+1}\right)\right)\right).$$

Solution by Chu Wenchang and Di Claudio Leontina Veliana, Universitàdegli Studi di Lecce, Lecce, Italy.

The sum is equal to 0. To prove this, we establish the more general result that for any real θ,

$$\sum_{k=1}^{n+1}(k+1)\sin^{k-1}(2\theta)\prod_{j=1}^{k}\left(\cot\theta - cot\frac{j\pi}{k+1}\right) = 2^{n+1}\frac{\sin(n+1)\theta}{\sin^2\theta}\cos^{n+1}\theta. \qquad (1)$$

Setting $\theta = m\pi/(n+1)$, we see that the sum in the problem statement is 0.

Because

$$\sin((k+1)\theta) = \operatorname{Im}\left((\cos\theta + i\sin\theta)^{k+1}\right) = \sum_{0\le j\le k/2}(-1)^j\binom{k+1}{2j+1}\sin^{2j+1}\theta\cos^{k-2j}\theta,$$

it follows that $\sin(k+1)\theta/\sin^{k+1}\theta$ is a polynomial of degree k in $\cot\theta$ with leading coefficient $k+1$. The polynomial takes the value 0 if and only if $\cot\theta = \cot(j\pi/(k+1))$, $1 \le j \le k$, so

$$\frac{\sin(k+1)\theta}{(k+1)\sin^{k+1}\theta} = \prod_{j=1}^{k}\left(\cot\theta - \cot\frac{j\pi}{k+1}\right).$$

Hence the left-hand side of (1) is equal to

$$\begin{aligned}\sum_{k=1}^{n+1}\sin^{k-1}(2\theta)\frac{\sin(k+1)\theta}{\sin^{k+1}\theta} &= \frac{1}{\sin^2\theta}\sum_{k=1}^{n+1}2^{k-1}\sin(k+1)\theta\cos^{k-1}\theta \\ &= \operatorname{Im}\left(\frac{e^{2i\theta}}{\sin^2\theta}\sum_{k=1}^{n+1}(2e^{i\theta}\cos\theta)^{k-1}\right) \\ &= \operatorname{Im}\left(\frac{e^{2i\theta}}{\sin^2\theta}\frac{1-2^{n+1}e^{(n+1)i\theta}\cos^{n+1}\theta}{1-2e^{i\theta}\cos\theta}\right) \\ &= -\frac{1}{\sin^2\theta}\operatorname{Im}\left(1-2^{n+1}e^{(n+1)i\theta}\cos^{n+1}\theta\right) \\ &= 2^{n+1}\frac{\sin(n+1)\theta}{\sin^2\theta}\cos^{n+1}\theta.\end{aligned}$$

This completes the proof of (1).

Also solved by Tsehaye Andebrhan, Daniele Donini (Italy), Ovidiu Furdui, Rolf Richberg (Germany), Michael Vowe (Switzerland), Li Zhou, and the proposer.

LCM Divisors **February 2003**

1664. *Proposed by Tim Ferguson, student, Linganore High School, Frederick, MD, and Lenny Jones, Shippensburg University, Shippensburg, PA.*

Given a positive integer n, a sequence $\lambda_1, \lambda_2, \ldots, \lambda_k$ of positive integers is called a partition of n if $\sum_{j=1}^{k}\lambda_j = n$. Given a partition $\pi : \lambda_1, \lambda_2, \ldots, \lambda_k$ of n, let $\mathrm{LCM}(\pi) = \mathrm{LCM}(\lambda_1, \lambda_2, \ldots, \lambda_k)$, and define

$$M_n = \max\left\{\mathrm{LCM}(\pi) : \pi \text{ a partition of } n\right\}.$$

Let p be a prime such that p^a divides M_n for some integer $a \ge 3$. Prove that if q is a prime with $p < q < p^{a-1}$, then q divides M_n.

Solution by Chip Curtis, Missouri Southern State University, Joplin, MO.

Fix n and let $\pi : \lambda_1, \lambda_2, \ldots \lambda_k$ be a partition of n for which $\mathrm{LCM}(\pi) = M_n$. Let p be a prime such that p^a divides M_n with $a \ge 3$. Without loss of generality we may assume that $p^a \mid \lambda_k$, so $\lambda_k = p^a t$ for some positive integer t. Let q be a prime with $p < q < p^{a-1}$, and suppose that q does not divide M_n. Then

$$p^{a-1}t + qt < 2p^{a-1}t \le p^a t.$$

Thus there is a positive integer $l > k$ so that the sequence $\mu_1, \mu_2, \ldots, \mu_l$ defined by

$$\mu_j = \begin{cases} \lambda_j & \text{if } 1 \le j \le k-1 \\ p^{a-1}t & \text{if } j = k \\ qt & \text{if } j = k+1 \\ 1 & \text{if } k+1 < j \le l \end{cases}$$

is a partition of n. But then

$$\begin{aligned}\mathrm{LCM}(\mu_1,\ \mu_2,\ldots,\mu_l) &= \mathrm{LCM}\left(\mathrm{LCM}(\lambda_1,\ldots\lambda_{k-1}),\ p^{a-1}t, qt\right)\\ &= \mathrm{LCM}\left(\mathrm{LCM}(\lambda_1,\ldots\lambda_{k-1}),\ p^{a-1}t, q\right)\\ &= q\cdot\mathrm{LCM}\left(\mathrm{LCM}(\lambda_1,\ldots,\lambda_{k-1}),\ p^{a-1}t\right)\\ &\geq q\frac{M_n}{p} > M_n,\end{aligned}$$

contradicting the definition of M_n.

Also solved by Roy Barbara (Lebanon), Jean Bogaert (Belgium), Con Amore Problem Group (Denmark), Daniele Donini (Italy), Robert L. Doucette, Kathleen E. Lewis, Reiner Martin, Bill Mixon, Rolf Richberg (Germany), Joel Schlosberg, Achilleas Sinefakopoulos, Li Zhou, and the proposers.

An Acute Inequality, Occasionally Obtuse **February 2004**

1665. *Proposed by Mihàly Bencze, Brasov, Romania.*

Let M be a point in the interior of triangle ABC and let P, Q, and R be the projections of M onto BC, CA, and AB, respectively. Prove that

$$MA^2\sin^2\frac{A}{2}+MB^2\sin^2\frac{B}{2}+MC^2\sin^2\frac{C}{2}\leq MP^2+MQ^2+MR^2.$$

Solution by Robert L. Doucette, McNeese State University, Lake Charles, LA.

The result is not true as stated. In all that follows,we assume that $A\leq B\leq C$. In order to better describe the conditions under which the result does hold, define the constant $C_0 = 2\arcsin((1+\sqrt{13})/6)\approx 100.28^\circ$, and the function $A_0:(\pi/2, C_0]\to\mathbb{R}$ by

$$A_0(C)=\frac{\pi-C}{2}-\arccos\left(-\frac{\sin(C/2)}{\sec C+3}+\sqrt{\left(\frac{\sin(C/2)}{\sec C+3}\right)^2+\frac{1+\cos C}{2}}\right). \quad (1)$$

We show that the inequality in the problem statement holds only in the following cases:

(i) if $C\leq\pi/2$

(ii) if $\pi/2 < C\leq C_0$ and $A_0(C)\leq A\leq(\pi-C)/2$.

Let $x = MP$, $y = MQ$, and $z = MR$. Referring to the inequality in the problem statement, let Δ be the expression obtained when the left-hand side is subtracted from the right-hand side. We first show that

$$4\Delta=\left(1-\tan^2\frac{A}{2}\right)(y-z)^2+\left(1-\tan^2\frac{B}{2}\right)(z-x)^2+\left(1-\tan^2\frac{C}{2}\right)(x-y)^2. \quad (2)$$

Note that this immediately establishes the result in case (i).

Because $\angle AQM$ and $\angle ARM$ are right angles, points A, R, M, and Q all lie on a circle with diameter MA. If M and A lie on opposite sides of $\overline{QR}$, then the sum of the measures of the arcs intercepted by $\angle A$ and $\angle QMR$ is 360°. If M and A are on the same side of $\overline{QR}$, then $\overline{QR}$ subtends both the supplement of $\angle A$ as well as $\angle QMR$. In either case, we conclude that $\angle QMR = 180^\circ-\angle A$. Applying the Extended Law of Sines to $\triangle QMR$, we obtain $QR = MA\sin A$. Applying the Law of Cosines yields $QR^2 = y^2+z^2+2yz\cos A$. Combining these two equations and using some standard trigonometric identities we obtain

$$MA^2 \sin^2 \frac{A}{2} = \frac{1}{4}\left(1 + \tan^2 \frac{A}{2}\right)(y-z)^2 + yz = \frac{1}{4}\tan^2 \frac{A}{2}(y-z)^2 + \frac{1}{4}(y+z)^2.$$

Using analogous formulas for $MB^2 \sin^2 \frac{B}{2}$ and $MC^2 \sin^2 \frac{C}{2}$, the left hand side of the inequality in the problem statement becomes

$$\frac{1}{4}\left(\tan^2 \frac{A}{2}(y-z)^2 + \tan^2 \frac{B}{2}(x-z)^2 + \tan^2 \frac{C}{2}(x-y)^2 + (y+z)^2 + (z+x)^2 + (x+y)^2\right). \tag{3}$$

The right-hand side, $x^2 + y^2 + z^2$, is equal to

$$\frac{1}{4}\left((y-z)^2 + (z-x)^2 + (x-y)^2 + (y+z)^2 + (z+x)^2 + (x+y)^2\right). \tag{4}$$

Subtracting (3) from (4) gives (2).

Now assume that $\pi/2 < C$, and for convenience, let $t(\alpha) = 1 - \tan^2(\alpha/2)$. With some algebra (including quadratic forms), we obtain

$$4\Delta = \lambda_1\big(\cos\theta(y-z) - \sin\theta(z-x)\big)^2 + \lambda_2\big(\sin\theta(y-z) + \cos\theta(z-x)\big)^2,$$

where $\theta = \frac{1}{2}\arctan(\frac{2t(C)}{t(B)-t(A)})$, and λ_1 and λ_2 correspond to the plus and minus sign choices, respectively, in

$$\frac{t(A) + t(B) + 2t(C) \pm \sqrt{\big(t(A) - t(B)\big)^2 + 4t(C)^2}}{2}.$$

If we fix C, we may consider λ_1 and λ_2 as functions of the single variable A. It is not difficult to show that $\lambda_1(A) > 0$ and $\lambda_2(A)$ is increasing for $0 < A \le (\pi - C)/2$.

It follows that $\mu(C) = \max\{\lambda_2(A) : 0 \le A \le (\pi - C)/2\} = t(\pi - C)/2 + 2t(C)$. The function μ is decreasing on $(\pi/2, \pi)$. Because $\lim_{C\to\pi/2^+} \mu(C) > 0$ and $\lim_{C\to\pi^-} \mu(C) = -\infty$, μ has a unique zero in the interval $(\pi/2, \pi)$. With a bit of effort, it can be shown that the value of this zero is C_0, the constant defined earlier. If $C_0 < C < \pi$, then $\lambda_2(A) < 0$ for all $0 < A \le (\pi - C)/2$.

Now consider the case in which $\pi/2 < C < C_0$. Because $\lambda_2(A)$ is an increasing function of A and $\lim_{A\to 0^+} \lambda_2(A) < 0$, there is exactly one $A \in (0, (\pi - C)/2)$ such that $\lambda_2(A) = 0$. Finding this requires some effort; the solution, $A_0 = A_0(C)$, is given in (1). If $C \in (\pi/2, C_0]$ and $A \in [A_0, (\pi - C)/2]$, then $\lambda_2 \ge 0$ and the desired inequality holds for any M in the triangle. If $C \in (\pi/2, C_0]$ and $A \in (0, A_0)$, then $\lambda_2(A) < 0$.

If $\lambda_2 < 0$, then we may exhibit a point M for which the desired inequality does not hold. First observe that $\theta \in (0, \pi/4]$. Let r be the inradius of $\triangle ABC$ and a, b, c the lengths of the sides opposite A, B, C respectively. There is a one-to-one correspondence between points M in the interior of $\triangle ABC$ and triples (x, y, z) of positive real numbers such that

$$ax + by + cz = (a + b + c)r. \tag{5}$$

To find a point M such that $\Delta < 0$, it suffices to find a positive triple $(x, y, z) \ne (r, r, r)$ satisfying (5) and with $\cos\theta(y - z) - \sin\theta(z - x) = 0$. We seek to determine appropriate values of k and z with $x = z - k\cos\theta$ and $y = z + k\sin\theta$. We then have

$$ax + by + cz - (a + b + c)r = (a + b + c)(z - r) + (b\sin\theta - a\cos\theta)k.$$

If $b\sin\theta - a\cos\theta = 0$, then let $z = r$ and $k = r/(2\cos\theta)$. If $b\sin\theta - a\cos\theta \neq 0$, then choose z sufficiently close to r so that

$$0 < \frac{a+b+c}{b\sin\theta - a\cos\theta}(r-z) < \frac{z}{\cos\theta}, \text{ and choose } k = \frac{a+b+c}{b\sin\theta - a\cos\theta}(r-z).$$

Note. A similar problem appeared as Problem 10970 in The American Mathematical Monthly, Vol. 109, No. 9, November, 2002.

Also solved by Herb Bailey, Michel Bataille (France), Knut Dale (Norway), Daniele Donini (Italy), Ovidiu Furdui, Julien Grivaux (France), John G. Heuver (Canada), Enkel Hysnelaj (Australia), Elias Lampakis (Greece), Murray S. Klamkin (Canada), Vivek Kumar Mehra (India), Peter E. Nüesch (Switzerland), Raul A. Simon (Chile), Helen Skala, Michael Vowe (Switzerland), Li Zhou, and the proposer.

Answers

Solutions to the Quickies from page 70.

A937. Consider the lattice points with positive coordinates under the graph of $x = n/y - 1/2$. For integer $1 \leq k \leq n$, the number of such points with first coordinate k is $\lfloor \frac{n}{k+1/2} \rfloor$. Summing, we find the total number of lattice points is $\sum_{k=1}^{n} \lfloor \frac{n}{k+1/2} \rfloor$. Next note that for integer $1 \leq k \leq n$, the number of positive integer lattice points with second coordinate k to the left of the curve is $\lfloor n/k - 1/2 \rfloor = \lfloor n/k + 1/2 \rfloor - 1$. Thus the total number of such points is $\sum_{k=1}^{n} \lfloor n/k + 1/2 \rfloor - n$. This completes the proof.

A938. Let $a,\ b \in G$ with $a \neq b$. Then $G = aG = bG$, so

$$s = \sum_{g\in G} g = \sum_{g\in G} ag = as = \sum_{g\in G} bg = bs.$$

It follows that $(a-b)s = 0$. Because $b - a \neq 0$, we conclude that either $s = 0$ or s is a zero divisor in R.

As an example for which $s \neq 0$, consider the ring $\mathbb{Z}/9\mathbb{Z}$ of integers modulo 9. Let $G = \{1, 4, 7\}$ be the cyclic subgroup under multiplication generated by 4. The sum of the elements of G is 3, a nonzero divisor of zero in $\mathbb{Z}/9\mathbb{Z}$.

For a second example let $M_3(F)$ denote the ring of 3×3 matrices over the field F, and let G be the multiplicative subgroup

$$G = \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \right\}.$$

The sum of the elements of G is

$$\begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix},$$

which is a nonzero zero divisor when F is not of characteristic 2. Note that both the ring and the group (which is isomorphic to the symmetric group S_3) are noncommutative.

REVIEWS

PAUL J. CAMPBELL, *Editor*
Beloit College

Assistant Editor: Eric S. Rosenthal, West Orange, NJ. Articles and books are selected for this section to call attention to interesting mathematical exposition that occurs outside the mainstream of mathematics literature. Readers are invited to suggest items for review to the editors.

Kolata, Gina, In Archimedes' puzzle, a new eureka moment, *New York Times* (14 December 2003) 1, 46 (National Edition, pp. 1, 32); `http://www.nytimes.com/2003/12/14/science/14MATH.html` .

Tangrams are seven polygons that form a square; children have fun rearranging the pieces into objects, animals, and more. Dictionaries cite the Chinese Tang dynasty (618–907) as the origin of the name. But a reinterpretation of Archimedes' *Stomachion* suggests his familiarity with such puzzles centuries earlier and an interest in combinatorial questions. Reviel Netz (Stanford University) interprets a diagram in the Archimedes palimpsest, with the word "multitude," as asking how many ways 14 polygons can form a square (answer: 17,152). We cannot be sure that this was Archimedes' intention, since the surviving fragment of the *Stomachion* contain nothing further on the topic. However, Netz's speculation—a far better explanation of the fragmentary *Stomachion* than any other—raises the prospect that Archimedes was a pioneer also in the field of combinatorics.

Brynsrud, Espen, Swede helps crack historic math problem, *Aftenposten Nettutgaven* (26 November 2003); `http://www.aftenposten.no/english/world/article.jhtml?articleID=678371` . Whitehouse, David, Historic maths problem 'cracked', BBC (27 November 2003), `http://news.bbc.co.uk/1/hi/sci/tech/3243736.stm` . Whitfield, John, Mathematicians dispute proof of century-old problem, *Nature* (9 December 2003), `http://www.nature.com/nsu/031208/031208-4.html` . Maths muddle, *New Scientist* (13 December 2003) 19. Roy, Edmond, 22-year-old cracks historic maths problem [includes interview], ABC (Australia) (20 December 2003), `http://www.abc.net.au/am/content/2003/s1014078.htm` . Oxenhielm, Elin, On the second part of Hilbert's 16th problem (author-corrected proof) (3 December 2003) `http://www.sciencedirect.com/` . Zhou, Yishao, Disclaimer `http://www.math.su.se/~yishao/16thproblem.shtml` . Discussion: `http://www.unstruct.org/archives/000186.html` . Oxenhielm, Elin, About ... the mathematical criticism on my paper (8 December 2003), `http://www.oxenhielm.com/` .

A Swedish graduate student, Elin Oxenhielm at Stockholm University, may in a few hours have solved the second part of Hilbert's 16th problem. That problem is about algebraic curves and surfaces; its second part, about boundary cycles for polynomial differential equations, is to show that the number of periodic solutions to a differential equation is finite. Her paper appears (for $30) only online from *Nonlinear Analysis*; no preprints at `http://xxx.lanl.gov/archive/math`. Yet her adviser Yishao Zhou has published a disclaimer that "the paper is incorrect... I could not imagine that the article would be accepted," and other mathematicians have also objected. Meanwhile, the journal, which had sent the paper to a single referee, has halted print publication and sent the paper to two more referees. Oxenhielm furnishes encouraging emails from Zhou but refuses further comment except to say that "the journal is responsible for [the paper's] correctness" (!)

Finch, Stephen R., *Mathematical Constants*, Cambridge University Press, 2003; xix + 602 pp, $95. ISBN 0–521–81805–2.

Mathematicians are familiar with a great many mathematical constants, but there are more constants in this book than you could ever have imagined, particularly in the seven chapters after the one entitled "Well-Known Constants." The treatment of each constant or family includes its own set of references.

Lang, Robert J., *Origami Design Secrets: Mathematical Methods for an Ancient Art*, A K Peters, 2003; viii + 585 pp, $48 (P). ISBN 1–56881–194–2.

This book is not a "step-by-step recipe book for design" of origami figures, but a collection of "codified mathematical and geometric techniques for developing a desired structure." The art of origami is beautifully illustrated on high-quality glossy paper. The author devotes a chapter to tree theory, the mathematics underlying the tree method of origami design formulated in earlier chapters, and includes a very extensive bibliography.

Devlin, Keith, 2003: Mathematicians face uncertainty, *Discover* 25 (1) (January 2004) 36.

"[M]athematicians finally had to agree that their prized notion of 'absolute proof' is an unattainable ideal...." Author Devlin cites proofs announced in 2003 that weren't (twin prime conjecture, Poincaré conjecture) and the indecision of referees (after five years) on Thomas Hales's proof of the Kepler conjecture about packing of spheres. He goes on to his distinction between the unfortunately named "right-wing" and "left-wing" definitions of proof and concludes that mathematicians sometimes should (because they have to) "settle for ... proof beyond a reasonable doubt." *Discover* magazine ranked this story 8th among 100 science stories of 2003.

Sudan, Madhu, Quick and dirty refereeing, *Science* 301 (29 August 2003) 1191–1192.

After the mudslinging surrounding Oxenhielm's paper and Devlin's attack on absolutism in mathematics, mathematicians may be somewhat relieved to learn that author Sudan (winner of the Nevanlinna Prize) and others have shown that "a proof can be written in a format that makes error detection very easy." After all, as Sudan remarks, "Proofs are by definition [supposed to be] easy to verify, whereas theorems in general are hard to prove." The method is a probabilistic procedure (there goes certainty!), and the new format is called a *probabilistically checkable proof* (PCP). Sudan shows how proving a particular theorem can be converted into an instance of the traveling salesperson problem; he then notes that PCPs are based on a similar transformation of the theorem and its proof into polynomials, accompanied by a "validity relation" operator. "Given a polynomial [the theorem], does there exist another polynomial [a proof] of pre-determined degree such that the validity operator maps the pair to the zero polynomial?" Mathematicians may not adopt PCPs—after all, as Sudan notes, mathematicians "look to proofs for providing insight and intuition"—but PCPs may provide a method to verify correct execution of computer programs.

Kolata, Gina, What is the most important problem in math today?, *New York Times* (11 November 2003) D13.

OK, before you read the next line, what's your guess? Given that there have been three popular books about it in the past year or so, it must be ... the Riemann Hypothesis.

Senn, Stephen, *Dicing with Death: Chance, Risk and Health*, Cambridge University Press; xii + 251 pp, $75, $28 (P). ISBN 0–521–83259–4, 0–521–54023–2.

This book about "biostatistics" is the most engrossing popular book on probability that I have read in a long time. The likelihood ratio arises on p. 8 already, hypothesis tests per se do not appear, the normal distribution rears up about three-quarters of the way through, and dozens of concepts in probability make natural appearances in memorable applied contexts, some involving risk and death.

NEWS AND LETTERS

62nd Annual William Lowell Putnam Mathematical Competition

Editor's Note: Additional solutions will be printed in the *Monthly* later in the year.

Problems

A1 Let n be a fixed positive integer. How many ways are there to write n as a sum of positive integers,

$$n = a_1 + a_2 + \cdots + a_k,$$

with k an arbitrary positive integer and $a_1 \le a_2 \le \cdots \le a_k \le a_k + 1$? For example, with $n = 4$, there are four ways: 4, $2+2$, $1+1+2$, $1+1+1+1$.

A2 Let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be nonnegative real numbers. Show that

$$(a_1 a_2 \cdots a_n)^{1/n} + (b_1 b_2 \cdots b_n)^{1/n} \le ((a_1 + b_1)(a_2 + b_2) \cdots (a_n + b_n))^{1/n}.$$

A3 Find the minimum value of

$$|\sin x + \cos x + \tan x + \cot x + \sec x + \csc x|$$

for real numbers x.

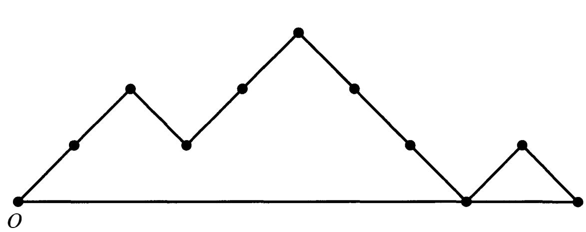
A4 Suppose that a, b, c, A, B, C are real numbers, $a \neq 0$ and $A \neq 0$, such that

$$|ax^2 + bx + c| \le |Ax^2 + Bx + C|$$

for all real numbers x. Show that $|b^2 - 4ac| \le |B^2 - 4AC|$.

A5 A Dyck n-path is a lattice path of n upsteps $(1, 1)$ and n downsteps $(1, -1)$ that starts at the origin O and never dips below the x-axis. A return is a maximal sequence of contiguous downsteps that terminates on the x-axis. For example, the Dyck 5-path illustrated has two returns, of length 3 and 1 respectively.

Show that there is a one-to-one correspondence between the Dyck n-paths with no return of even length and the Dyck $(n-1)$-paths.



A6 For a set S of nonnegative integers, let $r_S(n)$ denote the number of ordered pairs (s_1, s_2) such that $s_1 \in S$, $s_2 \in S$, $s_1 \neq s_2$, and $s_1 + s_2 = n$. Is it possible to partition the nonnegative integers into two sets A and B in such a way that $r_A(n) = r_B(n)$ for all n?

B1 Do there exist polynomials $a(x), b(x), c(y), d(y)$ such that

$$1 + xy + x^2y^2 = a(x)c(y) + b(x)d(y)$$

holds identically?

B2 Let n be a positive integer. Starting with the sequence $1, 1/2, 1/3, \ldots, 1/n$, form a new sequence of $n - 1$ entries $3/4, 5/12, \ldots, (2n-1)/(2n(n-1))$, by taking the averages of two consecutive entries in the first sequence. Repeat the averaging of neighbors on the second sequence to obtain a third sequence of $n - 2$ entries and continue until the final sequence produced consists of a single number x_n. Show that $x_n < 2/n$.

B3 Show that for each positive integer n,

$$n! = \prod_{i=1}^{n} \text{lcm}\{1, 2, \ldots \lfloor n/i \rfloor\}.$$

(Here lcm denotes the least common multiple, and $\lfloor x \rfloor$ denotes the greatest integer $\leq x$.)

B4 Let $f(z) = az^4 + bz^3 + cz^2 + dz + e = a(z - r_l)(z - r_2)(z - r_3)(z - r_4)$ where a, b, c, d, e are integers, $a \neq 0$. Show that if $r_1 + r_2$ is a rational number, and if $r_1 + r_2 \neq r_3 + r_4$, then r_1r_2 is a rational number.

B5 Let A, B, and C be equidistant points on the circumference of a circle of unit radius centered at O, and let P be any point in the circle's interior. Let a, b, c be the distances from P to A, B, C respectively. Show that there is a triangle with side lengths a, b, c, and that the area of this triangle depends only on the distance from P to O.

B6 Let $f(x)$ be a continuous real-valued function defined on the interval $[0, 1]$. Show that

$$\int_0^1 \int_0^1 |f(x) + f(y)| \, dx \, dy \geq \int_0^1 |f(x)| \, dx.$$

Solutions

Solution to A1 The answer is n. To see this, we will show that for each k, $1 \leq k \leq n$, there is a unique solution with k summands. Given a solution with k summands, there is an r between 0 and $k-1$ such that $n = a_1 + a_2 + \cdots + a_{k-r+1} + \cdots + a_k$ where $a_1 = \cdots = a_{k-r}$ and $a_{k-r+i} = \cdots = a_k = a_i + 1$. Thus, $n = a_1k + r$. This implies that $a_1 = \lfloor n/k \rfloor$, and r is the remainder of n mod k. So there can be only one such solution. But choosing the a_is as described provides one such solution.

Solution to A2 If any variable is 0 then the result is trivial, so we may assume that all are positive. Divide both sides by the right-hand side. Thus we are to show that

$$\left(\left(\frac{a_1}{a_1 + b_1}\right) \cdots \left(\frac{a_n}{a_n + b_n}\right)\right)^{1/n} + \left(\left(\frac{b_1}{a_1 + b_1}\right) \cdots \left(\frac{b_n}{a_n + b_n}\right)\right)^{1/n} \leq 1.$$

By two applications of the arithmetic-geometric mean inequality, we see that the left-hand side above is

$$\leq \frac{1}{n}\left(\frac{a_1}{a_1 + b_1} + \cdots + \frac{a_n}{a_n + b_n}\right) + \frac{1}{n}\left(\frac{b_1}{a_1 + b_1} + \cdots + \frac{b_n}{a_n + b_n}\right) = 1.$$

Solution to A3 Set $u = \sin x + \cos x$. Then $u^2 = 1 + 2\sin x \cos x$. Computation shows that the given expression is the absolute value of the function

$$u + \frac{u+1}{(u^2-1)/2} = u + \frac{2}{u-1} = 1 + (u-1) + \frac{2}{u-1} \equiv f(u).$$

When $u > 1$, applying the arithmetic-geometric mean inequality to the second and third terms of f shows that $f \geq 1 + 2\sqrt{2}$, with equality when $u = 1 + \sqrt{2}$. When $u < 1$, the same inequality shows that $f \leq 1 - 2\sqrt{2}$, with equality when $u = 1 - \sqrt{2}$. Taking absolute values, we see that the minimum value over the domain of interest is $2\sqrt{2} - 1$.

Solution to A4 Put $f(x) = ax^2 + bx + c$, $F(x) = Ax^2 + Bx + C$, $d = b^2 - 4ac$, $D = B^2 - 4AC$. By replacing f by $-f$ if necessary, we may assume $a > 0$. Similarly, we may assume $A > 0$.

If $D > 0$ then F has two distinct real roots r_1, r_2, and on taking $x = r_1$, $x = r_2$, we deduce that f also vanishes at the points r_i. Thus $f(x) = a(x - r_1)(x - r_2)$ and $F(x) = A(x - r_1)(x - r_2)$. The given inequality implies that $a \leq A$. Then $d = a^2(r_1 - r_2)^2 \leq A^2(r_1 - r_2)^2 = D$. If $D \leq 0$, we consider two cases:

Case 1. $d < 0$. On letting $x \to \infty$, $0 < a \leq A$. Since

$$\min_{x \text{ real}} F(x) = F\left(-\frac{B}{2A}\right) = -\frac{D}{4A} \quad \text{and} \quad \min_{x \text{ real}} f(x) = f\left(-\frac{b}{2a}\right) = -\frac{d}{4a},$$

it follows that

$$-\frac{D}{4A} = F\left(-\frac{B}{2A}\right) \geq f\left(-\frac{b}{2a}\right) = -\frac{d}{4a},$$

and so $0 \leq -d \leq -D$.

Case 2. $d \geq 0$. Since $F(x) \pm f(x) \geq 0$ for all x it follows that the discriminant of $F \pm f$ is ≤ 0. That is,

$$(B+b)^2 - 4(A+a)(C+c) \leq 0, (B-b)^2 - 4(A-a)(C-c) \leq 0.$$

On adding these two inequalities we find that $2(D + d) \leq 0$, so $0 \leq d \leq -D$.

Solution to A5 We exhibit a bijection between the two sets. Suppose we are given a Dyck n-path with no returns of even length. It begins with U and later returns to the x-axis. Now delete the path's first step U and the last step D of the first such return. The result is a Dyck $(n-1)$-path. This map is the desired bijection. To reverse it, suppose a Dyck $(n-1)$-path P is given; if P has no returns of even length, prepend UD to P, otherwise locate the initial segment of P through the *last* even-length return and "elevate" this segment, that is, put a U in front and a D after it.

Solution to A6 Yes. Let A be the set of nonnegative integers whose binary expansion has an even number of 1s and let B be those with an odd number of 1s. Given n and $a_1 \neq a_2 \in A$ with sum n, locate the first position in which the binary digits of a_1, a_2 differ (starting from the units digit) and interchange these digits. This gives a bijection from the ordered pairs counted by $r_A(n)$ to those counted by $r_B(n)$, with inverse given by the same procedure. So A, B form a partition as desired.

Solution to B1 There do not exist such polynomials. To see this, suppose there are such polynomials, and write $a(x) = a_0 + a_1x + a_2x^2 + \cdots$ and $b(x) = b_0 + b_1x + b_2x^2 + \cdots$. Then, equating coefficients of 1, x, and x^2, we would have the system of equations

$$1 = a_0 c(y) + b_0 d(y)$$
$$y = a_1 c(y) + b_1 d(y)$$
$$y^2 = a_2 c(y) + b_2 d(y).$$

This system has no solution because $c(y)$, $d(y)$ span at most a 2-dimensional subspace of the space of polynomials in y and $\{1, y, y^2\}$ would belong to this span, but these three polynomials are linearly independent.

Solution to B2 More generally, if we start with the sequence $a_1, a_2, \dots a_n$, we show inductively that the kth sequence is

$$\left\{\frac{1}{2^{k-1}}\sum_{r=0}^{k-1}\binom{k-1}{r}a_{i+r}\right\}_{1\le i\le n-k+1}$$

The base case $k = 1$ is trivial. Assume the result for k; then the ith entry in the $(k + 1)$st sequence is

$$\frac{1}{2}\left(\frac{1}{2^{k-1}}\sum_{r=0}^{k-1}\binom{k-1}{r}a_{i+r} + \frac{1}{2^{k-1}}\sum_{r=0}^{k-1}\binom{k-1}{r}a_{i+r+1}\right)$$
$$= \frac{1}{2^k}\left(\sum_{r=0}^{k-1}\binom{k-1}{r}a_{i+r} + \sum_{r=1}^{k-1}\binom{k-1}{r-1}a_{i+r}\right) = \frac{1}{2^k}\sum_{r=0}^{k}\binom{k}{r}a_{i+r}.$$

So, the final number x_n is

$$\frac{1}{2^{n-1}}\sum_{r=0}^{n-1}\binom{n-1}{r}a_{1+r} = \frac{1}{2^{n-1}}\sum_{r=0}^{n-1}\binom{n-1}{r}\frac{1}{1+r}$$
$$= \frac{1}{2^{n-1}}\sum_{r=0}^{n-1}\frac{1}{n}\binom{n}{r+1} = \frac{1}{n2^{n-1}}(2^n - 1) < \frac{2}{n}.$$

Solution to B3 For each prime p, we know that the power of p in $n!$ is $\sum_{k>0}\lfloor n/p^k\rfloor$. Thus it suffices to prove that the power of p on the right-hand side is the same. The power of p in $\operatorname{lcm}\{1, 2, ..., m\}$ is p^j where $p^j \le m < p^{j+1}$. Thus if k is given then the power of p in $\{1, 2, \dots, \lfloor n/i\rfloor\}$ is precisely p^k if $n/p^{k+1} < i \le n/p^k$. There are exactly $\lfloor n/p^{k+1}\rfloor - \lfloor n/p^k\rfloor$ such i. Hence the power of p in the right-hand side above is

$$\sum_k k\left(\left\lfloor\frac{n}{p^{k+1}}\right\rfloor - \left\lfloor\frac{n}{p^k}\right\rfloor\right) = \sum_{k=1}^{\infty}(k-1)\left\lfloor\frac{n}{p^k}\right\rfloor = \sum_{k=1}^{\infty}\left\lfloor\frac{n}{p^k}\right\rfloor.$$

Solution to B4 Clearly $f(r_1 + r_2)$ is a rational number. By the factored form of f we see that

$$f(r_1 + r_2) = ar_1r_2(r_1 + r_2 - r_3)(r_1 + r_2 - r_4)$$
$$= ar_1r_2(r_1 + r_2)(r_1 + r_2 - r_3 - r_4) + ar_1r_2r_3r_4.$$

Here $ar_1r_2r_3r_4 = e$ is an integer, and

$$r_1 + r_2 - r_3 - r_4 = 2(r_1 + r_2) - (r_1 + r_2 + r_3 + r_4) = 2(r_1 + r_2) + b/a$$

is rational. Hence

$$\frac{f(r_1+r_2)-e}{a(r_1+r_2-r_3-r_4)}=r_1r_2(r_1+r_2)$$

is rational. If $r_1+r_2\neq 0$ then we are done. If $r_2=-r_1$ then

$$ar_1^4+br_1^3+cr_1^2+dr_1+e=0=ar_1^4-br_1^3+cr_1^2-dr_1-e,$$

which gives $(br_1^2+d)r_1=0$. If $r_1=0$ then $r_1r_2=0$, a rational number, and we are done. Note that $b\neq 0$, since $b=0$ and $r_1+r_2=0$ together imply that $r_1+r_2=r_3+r_4$, contrary to our hypothesis. Thus $r_1=\pm\sqrt{-d/b}$ and $r_2=\mp\sqrt{-d/b}$, so that $r_1r_2=d/b$ is rational in this case also. This completes the proof.

Solution to B5 Representing points in the plane by complex numbers, we may take $A=1$, $B=\omega$, $C=\omega^2$, where $\omega=(-1+i\sqrt{3})/2$ is a cube root of unity. The line segments $\overline{AP}$, $\overline{BP}$, and $\overline{CP}$ then have lengths $|P-1|$, $|P-\omega|$, and $|P-\omega^2|$, which form the sides of a triangle if and only if there exist complex numbers z_1, z_2, z_3 (the triangle vertices) such that $|z_1-z_3|$, $|z_2-z_3|$, and $|z_3-z_2|$ are equal to $|P-1|$, $|P-\omega|$, and $|P-\omega^2|$.

Such numbers do exist, defined by $z_1-z_3=P-1$, $z_2-z_1=w(P-w)$, $z_3-z_2=w^2(P-\omega^2)$. The sum of these three complex numbers is zero, so, when considered as vectors, they are the sides of a triangle.

Write $P=x+iy$. The area of the triangle, found by computing the cross product of two of the sides, is (up to sign)equal to

$$\frac{1}{4}\left((x-1)\left(\frac{\sqrt{3}}{2}x-\frac{1}{2}y+\frac{\sqrt{3}}{2}\right)-y\left(-\frac{1}{2}x-\frac{\sqrt{3}}{2}y+\frac{1}{2}\right)\right)=\frac{\sqrt{3}}{2}(x^2+y^2-1).$$

But $x^2+y^2<1$, since P is inside the circle, and so the area of the triangle is given by $\sqrt{3}(1-r^2)/4$, where r is the distance from P to O.

Solution to B6 Let $\mathcal{P}=\{x\in[0,1]: f(x)\geq 0\}$ and $\mathcal{N}=\{x\in[0,1]: f(x)<0\}$. Put $P=\mu(\mathcal{P})$, $N=\mu(\mathcal{N})$ (the measures of $\mathcal{P}$ and $\mathcal{N}$, respectively). If either $P=0$ or $N=0$ then the inequality is obvious. Thus we may assume $P>0$ and $N>0$. Consider the average values of $|f|$ on $\mathcal{P}$ and $\mathcal{N}$:

$$\mu_p=\frac{1}{P}\int_{\mathcal{P}}f(x)\,dx,\qquad \mu_n=-\frac{1}{N}\int_{\mathcal{N}}f(x)\,dx.$$

By replacing $f(x)$ by $-f(x)$ if necessary, we may assume $\mu_p\geq\mu_n$. Clearly

$$\iint_{\mathcal{P}\times\mathcal{P}}|f(x)+f(y)|\,dx\,dy=2P^2\mu_p,\qquad \iint_{\mathcal{N}\times\mathcal{N}}|f(x)+f(y)|\,dx\,dy=2N^2\mu_n.$$

In addition,

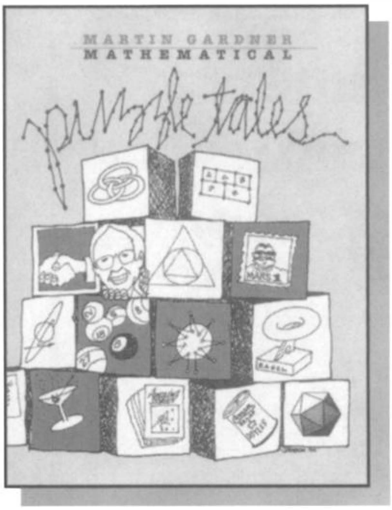
$$\iint_{\mathcal{P}\times\mathcal{N}}|f(x)+f(y)|\,dx\,dy\geq\left|\iint_{\mathcal{P}\times\mathcal{N}}(f(x)+f(y))\,dx\,dy\right|$$
$$=|NP\mu_p-NP\mu_n|=NP(\mu_p-\mu_n),$$

and the same inequality holds for the integral over $\mathcal{N}\times\mathcal{P}$. Thus, the given left-hand side is greater than or equal to

$$2P^2\mu_p+2NP(\mu_p-\mu_n)+2N^2\mu_n=P(\mu_p-\mu_n)+(2N-1)^2\mu_n+P\mu_p+N\mu_n$$
$$\geq P\mu_p+N\mu_n=\int_0^1|f(x)|\,dx.$$

Thanks to Byron Walden for editorial assistance.

The Mathematical Association of America



Mathematical Puzzle Tales

Martin Gardner

Series: Spectrum

Martin offers everybody (not just mathematicians) creative refuge for the imagination. The puzzles in this book are not just puzzles. Very often, they embody deep mathematical principals that deal with matters not yet well enough understood to be applied to the practical world. Such "games" are not more trivial than "real" mathematics. They may well be more important and may be the foreshadowing of future mathematics.
—Isaac Asimov from the Preface

Martin Gardner published his first book in 1935. Since then he has published more than 60 books, most of them about mathematics and sciences, but also philosophy and literature. He has charmed readers of all ages with his mathematical insights and sense of fun.

The MAA is proud to reissue this collection of thirty-six stories taken from Isaac Asimov's Science Fiction Magazine. Brilliant, amusing, these brainteasers will help you sharpen your wits and prepare for takeoff into uncharted universes of the future. The challenging problems presented here are based on geometry, logarithms, topology, probability, weird number sequences, logic and virtually every other aspect of mathematics as well as wordplay.

Included are: Lost on Capra • Space Pool • Machismo on Byronia • The Third Dr. Moreau • The Voyage of the Bagel • The Great Ring of Neptune • The Postage Stamps of Philo Tate • Captain Tittebaum's Tests • The Three Robots of Professor Tinker • How Bagson Bagged a Board Game • The Explosion of Blabbage's Oracle • No Vacancy at the Aleph-Null Inn...and more.

Catalog Code: MPT/JR 168 pp., Paperbound, 2001 ISBN 088385-533-x List: $22.50 MAA Member: $18.00

Name__________________ Credit Card No.__________________

Address__________________ Signature__________________Exp. Date____/____

City__________________ Qty______ Price $________ Amount $________

State________ Zip ________ Shipping and Handling $________

Phone __________________ **Catalog Code: MPT/JR** Total $________

Shipping and Handling: USA orders (shipped via UPS): $3.00 for the first book, and $1.00 for each additional book. **Canadian orders:** $4.50 for the first book and $1.50 for each additional book. Canadian orders will be shipped within 2-3 weeks of receipt of order via the fastest available route. We do not ship via UPS into Canada unless the customer specially requests this service. Canadian customers who request UPS shipment will be billed an additional 7% of their total order. **Overseas Orders:** $4.50 per item ordered for books sent surface mail. Airmail service is available at a rate of $10.00 per book. Foreign orders must be paid in US dollars through a US bank or through a New York clearinghouse. **Credit card orders are accepted for all customers. All orders must be prepaid with the exception of books purchased for resale by bookstores and wholesalers.**

Order Via:

Phone: **1 (800) 331.1622**
Fax: **(301) 206.9789**
Mail: **Mathematical Association of America**
PO Box 91112
Washington, DC 20090-1112
Web: **www.maa.org**

CONTENTS

THE MATHEMATICAL ASSOCIATION OF AMERICA
1529 Eighteenth Street, NW
Washington, DC 20036